

ARBEITSBERICHTE DES INSTITUTS FÜR MATHEMATISCHE MASCHINEN UND DATENVERARBEITUNG (INFORMATIK)

FRIEDRICH ALEXANDER UNIVERSITÄT ERLANGEN NÜRNBERG

**HERAUSGEBER : H. Billing, W. Händler, U. Herzog,
F. Hofmann, K. Leeb, P. Mertens, H. Müller, G. Nees,
H. Niemann, D. Seitzer, H. J. Schneider, H. Wedekind**

KLASSIFIKATORBEZOGENE MERKMALSAUSWAHL

Dieter Jäpel

Dissertation

BAND 13 · NUMMER 4 · ERLANGEN · JULI 1980

ARBEITSBERICHTE DES INSTITUTS FÜR MATHEMATISCHE MASCHINEN UND DATENVERARBEITUNG (INFORMATIK)

FRIEDRICH ALEXANDER UNIVERSITÄT ERLANGEN NÜRNBERG

**HERAUSGEBER : H. Billing, W. Händler, U. Herzog,
F. Hofmann, K. Leeb, P. Mertens, H. Müller, G. Nees,
H. Niemann, D. Seitzer, H.J. Schneider, H. Wedekind**

KLASSIFIKATORBEZOGENE MERKMALSAUSWAHL

Dem Fachbereich Ingenieurwissenschaften
der Universität Erlangen-Nürnberg
zur Erlangung des Grades
DOKTOR - INGENIEUR

vorgelegt von
Dieter Jäpel

Erlangen, 1980

Dissertation



Tag der Einreichung: 29. 4. 1980

Tag der Promotion: 10. 7. 1980

Dekan: Prof. Dr. H. Niemann

Berichterstatter: Prof. Dr. H. Niemann

Prof. Dr. G. Nees

Zusammenfassung

Der Entwurf eines Systems zur Klassifikation von Mustern zerfällt klassischerweise in zwei Teile: den Entwurf einer Merkmalsextraktionsstufe und den Entwurf der eigentlichen Klassifikationsstufe. Diese Aufteilung hat ihren Grund darin, daß Klassifikatoren in der Regel nur relativ wenige Merkmale verarbeiten Können. In der Merkmalsauswahlstufe wird deshalb eine einfache Reduktion der Zahl der Merkmale angestrebt.

Im Rahmen dieser Arbeit wird das Konzept "Klassifikator-bezogene Merkmalsauswahl" eingeführt und experimentell erprobt. Diese Konzept fordert, daß der Entwurf der Merkmalsextraktionsstufe im Hinblick auf den daran anschließend zu verwendenden Klassifikator zu erfolgen hat. Die Experimente zeigen, daß das neue Konzept herkömmlichen Merkmalsauswahlverfahren in Bezug auf die Leistung des Gesamtsystems überlegen ist.

Abstract

The design of a classification system is normally decomposed into two distinct parts: the design of a feature extractor and the design of a classification procedure. The reason for this decomposition is that classical classification procedures can work efficiently only with relatively small sets of features. Feature extraction is a method to achieve simple reductions in the number of features.

This dissertation introduces the concept of "Classifier-determined Feature Selection". The concept states that the design of a feature extractor has to be done with respect to the structure of the classifier one has in mind to use. Numerical results show that the concept is superior to classical feature selection methods.

Inhaltsverzeichnis

1.	Einführung	1
2.	Klassifikation: die klassische Zweiteilung	9
2.1	Klassifikationsverfahren	10
2.1.1	Der optimale Bayesklassifikator	11
2.1.2	Einige Spezialisierungen des Bayes- klassifikators	15
2.1.2.1	Normalverteilungsklassifikator	16
2.1.2.2	Abstandmessende Klassifikatoren	17
2.1.3	Der im quadratischen Mittel angepaßte Klassifikator	20
2.1.4	Hinweise auf weitere Klassifikatortypen	22
2.2	Merkmalsbewertung und Merkmalsauswahl	22
2.2.1	Bewertungsmaße für Merkmale	23
2.2.2	Verfahren zur Merkmalsauswahl	27
2.2.2.1	Auswahlverfahren	28
2.2.2.2	Merkmalsauswahl durch Transformation	28
2.2.2.3	Karhunen-Loève-Transformation	30
3.	Klassifikatorbezogene Merkmalsauswahl	33
3.1	Das Konzept	33
3.2	Klassifikatorbezogene Merkmalsauswahl beim Bayesklassifikator	38
3.3	Klassifikatorbezogene Merkmalsauswahl beim modifizierten Minimum-Abstand- Klassifikator	43
3.4	Verallgemeinerung des Konzepts für andere Klassifikatoren	51

4.	Experimentelle Untersuchungen	53
4.1	Probleme bei der Realisierung von Musterkennungssystemen	55
4.2	Datensätze	59
4.3	Das Experimentiersystem	62
4.4	Zur Implementierung der Merkmalsauswahl- verfahren	65
4.4.1	Verfahren zur numerischen Optimierung	65
4.4.2	Besonderheiten bei der Bestimmung linearer Merkmalsauswahlverfahren	69
4.4.3	Realisierung der Auswahlverfahren	72
4.4.3.1	DIV-optimale Merkmale	73
4.4.3.2	MMA-optimale Merkmale	75
4.4.3.3	NVK-optimale Merkmale	83
4.5	Einige Resultate	85
5.	Zusammenfassung	95
	Literaturverzeichnis	99
Anhang A	Werte der Gütemaße bei der Berechnung der Transformationsmatrizen	107
Anhang B	Klassifikationsergebnisse bei Verwendung der KL-Koeffizienten als Merkmale	115
Anhang C	Ergebnisse der Klassifikation nach Anwen- dung der berechneten linearen Transfor- mation, geordnet nach Zieldimension	119
Anhang D	Ergebnisse der Klassifikation nach Anwen- dung der berechneten linearen Transfor- mation, geordnet nach Klassifikatoren	127
Anhang E	Klassenäquivalente Ellipsen für die Reduktionen von 30 auf 2 Merkmale	135
Anhang F	Verzeichnis der wichtigsten Symbole	147

1. Einführung

Die vorliegende Arbeit beschäftigt sich mit dem Thema Merkmalsauswahl, ein Thema, das einen Spezialbereich innerhalb der relativ jungen Wissenschaft "Mustererkennung" abdeckt. Bei dem Fach Mustererkennung handelt es sich um den Versuch, die Fähigkeit von Lebewesen, Objekte in ihrer Umgebung wahrzunehmen, mit ingenieurmäßigen Hilfsmitteln konstruktiv zu simulieren. Es wird wohlgemerkt nicht verlangt, daß Verfahrensweisen, wie sie Lebewesen anwenden, nachgebildet werden; es geht rein darum, Objekte mit Hilfe technischer Einrichtungen wahrzunehmen und - was für Lebewesen in diesem Zusammenhang selbstverständlich ist - zu identifizieren oder zu beschreiben.

Im wesentlichen wird mit der Mustererkennung also der Versuch unternommen, intellektuelle Fähigkeiten der Lebewesen und insbesondere des Menschen durch Maschinen zu ersetzen. Es darf in diesem Zusammenhang daran erinnert werden, daß ein ähnlicher Vorgang mit der Ersetzung der Muskelkraft durch Maschinen bereits abgelaufen ist. Die Ersetzung der intellektuellen Fähigkeiten des Menschen durch Maschinen ist jedoch eine ungleich schwierigere Aufgabe, da bis heute die Gesetzmäßigkeiten der menschlichen Wahrnehmung und des Denkens fast gänzlich unerforscht sind.

So bleibt der Mustererkennung als ingenieurmäßig betriebene Disziplin nur, einige exakt umrissene Aufgaben aus dem Bereich der menschlichen Wahrnehmung herauszugreifen und in nur loser Anlehnung an die Arbeitsweise menschlicher Gehirne Verfahren und Geräte zur Lösung solcher Aufgaben zu entwerfen. Es bleibt allenfalls ein irgendwie gearteter Leistungsvergleich zwischen Mensch und Maschine durchzuführen, welcher die Leistungsfähigkeit der entworfenen Methoden und Geräte in bezug zu dem setzt, was der Mensch zu leisten in der Lage ist.

Als Beispiele für solche Teilaufgaben seien die Fähigkeiten des Menschen, Sprache zu verstehen, Bücher zu lesen, Fahrzeugtypen zu unterscheiden, die Art eines Baumes "mit einem Blick" zu erkennen oder Bilder zu beschreiben, aufgeführt. Es handelt sich dabei zum einen um die Zusammenfassung verschiedener Objekte unter einem gemeinsamen Gesichtspunkt, wie zum Beispiel die Erkennung eines Baumes als eine Birke, zum andern um die Analyse von Objekten bezüglich ihres Aufbaues und einer eventuell nachfolgenden Beschreibung des Gesamtobjektes wie bei der Beschreibung eines Bildes von einem Zimmer. Die Mustererkennung kann dementsprechend auch in zwei Bereiche unterteilt werden, die im folgenden kurz beschrieben werden sollen.

Der erste Bereich wird mit der Phrase "Klassifikation einfacher Muster" umrissen. Die Grundaufgabe in diesem Bereich ist die Zuordnung eines Objektes, das Muster genannt wird, zu einer Menge von Objekten mit ähnlichen Eigenschaften. Diese Menge von Objekten heißt Klasse oder Musterklasse. Beim Beispiel von den Bäumen wären also die Muster Bäume, die Klassen Baumarten. Das Grundproblem ist die Zuordnung eines Baumes zu einer Art. Dieses Problem heißt das Klassifikationsproblem.

Als zweiter Bereich wird der Bereich "Analyse komplexer Muster" in der Mustererkennung behandelt. Die Grundaufgabe in diesem Bereich ist die Beschreibung eines Objektes, das dann neben "Muster" oft die Bezeichnung Szene erhält, in seinem Aufbau aus einfachen Bestandteilen, die in diesem Zusammenhang als Grundsymbole bezeichnet werden. In dem einfachen Beispiel von den Bäumen könnte eine Szene durch eine Baumgruppe gebildet werden. Aufgabe der Szenenanalyse wäre zum Beispiel, Anzahl, Art und Position der Bäume in der Szene zu beschreiben. Das Ergebnis ist eine Bildbeschreibung.

Es darf natürlich nicht übersehen werden, daß die hier ange-deuteten Bereiche "Klassifikation einfacher Muster" und

"Analyse komplexer Muster" in bezug auf eine konkrete Problemstellung zu sehen sind. Das angestrebte Ziel entscheidet letztlich darüber, ob ein Objekt als komplexes oder als einfaches Muster angesehen wird: Ein Waldstück kann zu dem Zweck betrachtet werden, festzustellen, ob es sich um Laubwald oder Nadelwald handelt. In diesem Fall liegt ein Klassifikationsproblem vor und das Muster "Waldstück" wird als einfaches Muster behandelt. Das gleiche Waldstück kann aber auch mit dem Ziel betrachtet werden, seinen Aufbau aus Einzelbäumen zu beschreiben. Dann wird das gleiche Erscheinungsbild als komplexes Muster betrachtet werden müssen.

Es muß auch berücksichtigt werden, daß sich beide Teilbereiche der Mustererkennung gegenseitig als Hilfsmittel zur Lösung der eigenen Aufgabenstellung heranziehen. So kann es notwendig sein, ein Waldstück erst in seinem Aufbau zu analysieren, ehe es als Laub- oder Nadelwald klassifiziert werden kann. Andererseits mag eine Beschreibung eines Waldstückes einen Klassifikationstest benötigen, um die Arten der vorkommenden Bäume festzustellen. Die Unterscheidung zwischen beiden Bereichen der Mustererkennung ist an dieser Stelle sicherlich unscharf.

Die vorliegende Arbeit beschäftigt sich mit einem Problem der Mustererkennung, das dem Bereich Klassifikation zuzuordnen ist. Es geht dabei um die Frage, welche Kriterien dazu verwendet werden können, Muster mit einer unbekannten Klassenzugehörigkeit einer bekannten Klasse von Mustern zuzuordnen. Im Zusammenhang mit dem Klassifikationsproblem bezeichnet man solche Kriterien als Merkmale. Davon zu unterscheiden sind die sogenannten Klassencharakteristika, die dem Klassifikator als Beschreibung der Musterklassen zugeführt werden müssen. Der Klassifikator entscheidet über die Klassenzugehörigkeit von Mustern aufgrund der ihm bekannten Klassencharakteristika und der ihm angebotenen Merkmale.

Es ist nun häufig der Fall, daß es eine Vielzahl von Eigenschaften von Mustern gibt, die als Kriterien für die Klassifikation herangezogen werden können. Nicht alle diese Kriterien sind jedoch gleich gut als Merkmale für die Klassifikation geeignet. Es stellt sich also das Problem, gute Merkmale für die Klassifikation zu finden, speziell in dem Fall, wo man viele Kandidaten für Merkmale kennt. Es handelt sich dann um ein Auswahlproblem, das Problem der Merkmalsauswahl oder das Problem der Merkmalsextraktion. Einige Teilaspekte genau dieses Problems behandelt die vorliegende Arbeit.

Es ist mit den bisher vorgestellten Begriffen nun möglich, den prinzipiellen Aufbau eines Klassifikationssystems, also eines Systems zur automatischen Klassifikation von Mustern, darzustellen. Der schematische Aufbau ist in Bild 1-1 dargestellt.

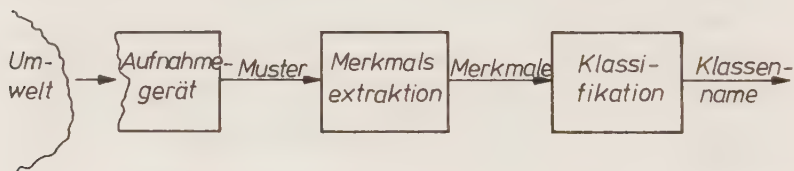


Bild 1-1 Aufbau eines Klassifikationssystems

Ein Aufnahmegerät, z. B. eine Kamera oder ein Mikrofon, bringt den interessierenden Teil der Umwelt, das Muster, in eine maschinengerechte Darstellung, eine Merkmalsextraktionsstufe extrahiert die interessierenden Merkmale, die nun einem Klassifikator angeboten werden, der für die Entscheidung über die Klassenzugehörigkeit zuständig ist. Die so in Bild 1-1 angedeutete Aufteilung in Merkmalsextraktion und Klassifikation wird in der Mustererkennungsliteratur klassischerweise durchgehalten.

In der vorliegenden Arbeit wird für die Realisierung eines

Mustererkennungssysteme nach Bild 1-1 die Gruppe von Verfahren zugrunde gelegt, die allgemein als statistische Methoden der Mustererkennung bezeichnet werden. Hierfür wird angenommen, daß die für die Entscheidung relevanten Merkmale Zahlen sind, die zu einem Vektor zusammengefaßt werden können. Dieser Merkmalsvektor wird aufgefaßt als eine Zufallsvariable. Die Beobachtung eines Musters ist somit in dem Modell die Beobachtung der Zufallsvariablen. Die Klassifikation stellt sich damit als die Berechnung von Wahrscheinlichkeiten für das Vorliegen einer speziellen Klasse dar. Die Klassencharakteristika werden bei diesem Ansatz in Form von Wahrscheinlichkeitsverteilungen für den Merkmalsvektor ausgedrückt. Die jetzt statistische Aufgabe der optimalen Klassifikation von Mustern mit statistischen Hilfsmitteln hat in der Mustererkennung starke Beachtung gefunden und zur Entwicklung einer umfassenden Theorie geführt, der der statistischen Mustererkennung /F1/. Diese Theorie existiert zunächst ohne jede Bezugnahme auf eine Merkmalsextraktionsstufe.

Im Zusammenhang mit statistischer Klassifikation ist die Aufgabe der Merkmalsextraktionsstufe die Bereitstellung solcher Merkmale, die günstige Eigenschaften für die Klassifikation aufweisen. Das Problem stellt sich häufig so dar, daß aus einer relativ großen Menge von Meßwerten am vorliegenden Muster eine relativ kleine Zahl von Merkmalen gebildet werden soll. Wie dies zu geschehen hat, ist in der Entwurfsphase für die Merkmalsextraktionsstufe festzulegen. Ein mögliches Konzept für den Entwurf einer Merkmalsextraktionsstufe stellt diese Arbeit vor. Es wird hierbei mit der Annahme gearbeitet, daß bereits das Muster mit Hilfe eines Zufallsvektors beschrieben wird. Gesucht ist lediglich eine Abbildung, die die große Komponentenzahl im "Mustervektor" auf eine kleine Komponentenzahl im "Merkmalsvektor" abbildet.

Ein wie oben beschriebenes Konzept für Klassifikation und Merkmalsauswahl wird in verschiedenen Anwendungsbereichen

der Mustererkennung verwendet, um ein konkretes Erkennungssystem zu erhalten. Ein sehr bekanntes Anwendungsgebiet ist die Klassifikation von Schriftzeichen /S1/. Weitere Anwendungen sind die Klassifikation isoliert gesprochener Worte (Befehlswortererkennung), die Zuteilung von EEG's zu Schlafzuständen, die Einteilung von EKG's in normal und pathologisch und andere. Eine Übersicht gibt /N1/.

Damit kann der Ausgangspunkt der vorliegenden Arbeit wie folgt umrissen werden: Betrachtet wird die Anpassung einer Merkmalsextraktionsstufe an eine Klassifikationsstufe. Dabei liegt ein statistisches Modell für die Musterklassen zugrunde, die Klassifikation greift auf statistische Methoden zurück. Es wird weiterhin angenommen, daß die statistischen Eigenschaften der Musterklassen sich nicht ändern, es ist also keine fortlaufende Anpassung des Systems an veränderliche Umgebungsbedingungen notwendig. Man verwendet hier das Schlagwort "fest dimensionierte, statistische Klassifikatoren" /N2/.

Das Anliegen der vorliegenden Arbeit ist, ein Verfahren zum Entwurf einer Merkmalsauswahlstufe anzugeben. Dabei wird erstmals versucht, die Merkmalsauswahlstufe so zu gestalten, daß die Struktur der nachfolgenden Klassifikationsstufe in die Entwurfskriterien mit eingeht. Das resultierende Prinzip wird mit "klassifikatorbezogener Merkmalsauswahl" bezeichnet. Die Tragfähigkeit des Konzeptes soll entgegen den bisherigen Gepflogenheiten in der Literatur an Datensätzen aus grundsätzlich verschiedenen Problemkreisen aufgezeigt werden, so daß ein gewisser Grad von Abstraktion vom Einzelproblem gewährleistet wird.

Die vorliegende Arbeit stellt - nach einer Darstellung des klassischen Ansatzes mit Klassifikationstheorie und Merkmalsauswahltheorie im zweiten Kapitel - das Konzept der klassifikatorbezogenen Merkmalsauswahl im dritten Kapitel dar. Im dritten Kapitel wird das Konzept auch exemplarisch

an zwei speziellen Klassifikatoren vorgeführt. Das vierte Kapitel enthält daran anschließend die Beschreibung der mit Hilfe des Konzeptes "klassifikatorbezogene Merkmalsauswahl" durchgeführten Experimente sowie den Vergleich zu einem klassisch gewordenen Merkmalsauswahlverfahren. Im fünften und letzten Kapitel wird schließlich noch eine Zusammenfassung der Resultate gegeben.

2. Klassifikation: die klassische Zweiteilung

In der vorliegenden Arbeit wird folgendes Modell für das Klassifikationsproblem verwendet: Gegeben sind k Klassen, die stets mit dem Namen Ω_k bezeichnet werden. Jedes Muster, bezeichnet als $\rho_{\underline{f}}$, wird durch die Merkmalsextraktionsstufe zunächst auf einen n -komponentigen Zahlenvektor abgebildet, so daß man ein Muster geometrisch als einen Punkt im n -dimensionalen euklidischen Raum, dem $\mathbb{R}^n$, auffassen kann. Dieser Merkmalsvektor $\rho_{\underline{c}}$ wird einer Entscheidungsstufe, dem Klassifikator, angeboten, der aus der Kenntnis von $\rho_{\underline{c}}$ und den Klassencharakteristika die Zugehörigkeit von $\rho_{\underline{f}}$ zu einer der k Klassen Ω_k ableitet.

Es ist in einigen Anwendungsfällen möglich, so viele Merkmale aus einem Muster $\rho_{\underline{f}}$ zu erzeugen, daß die aus den üblichen statistischen Ansätzen gewonnenen Klassifikatoren nicht mehr mit vertretbarem Aufwand realisiert werden können. So ist es zum Beispiel durchaus üblich, Schriftzeichen als Matrizen mit 30×40 binären Elementen darzustellen, was 1200 Meßwerten am Muster bzw. einem 1200-dimensionalen Merkmalvektor entspricht. Ein Klassifikator, der von der Theorie her alle Merkmale und alle möglichen Produkte aus zwei Merkmalen mit Gewichten versieht, müßte demnach für jede der betrachteten Klassen 721.800 Koeffizienten speichern und für jedes zu klassifizierende Schriftzeichen verarbeiten. In einem 36-Klassen-Problem, was in etwa dem Zeichensatz von FORTRAN-Programmen entspricht, sind dies über 25 Millionen Koeffizienten; selbst mit heutiger Schaltkreistechnik ein gigantisches Rechen- und Speicherungsproblem.

Es ist deshalb erforderlich, n -dimensionale Merkmalvektoren in ihrer Dimension zu reduzieren. Bei dieser Dimensionsreduktion sollte möglichst wenig Information über die Klassenzugehörigkeit der Muster verloren gehen. Weiterhin strebt man an, diese Dimensionsreduktion möglichst unabhängig von der Klassifikation zu sehen. Man entwickelte zwei getrennte

Theorien innerhalb der Mustererkennung: eine Theorie der Klassifikation und eine Theorie der Dimensionsreduktion, die in diesem Zusammenhang häufig auch als Merkmalsextraktion oder als Merkmalsauswahl bezeichnet wird. Im Rahmen dieser Arbeit soll das Wort Merkmalsauswahl für die Theorie der Reduktion der Dimension bei Merkmalvektoren stehen.

Gemäß dieser, auch in der Mustererkennungsliteratur /N1, F1, T2, Y1/ durchgehaltenen Trennung, wird der klassische Ansatz im folgenden in zwei Abschnitten dargestellt. Abschnitt 2.1 beschäftigt sich mit dem abstrakten Problem Klassifikation, das heißt, es wird dargestellt, in welcher Weise ein n -dimensionaler Merkmalvektor $\underline{c}$ zu einer von k Klassen Ω_k zugeordnet werden kann. Abschnitt 2.2 dagegen stellt die Methoden vor, mit deren Hilfe ein n -dimensionaler Merkmalsvektor $\underline{c}$ in einen n' -dimensionalen Merkmalsvektor $\underline{c}'$ überführt werden kann mit der Maßgabe, daß möglichst wenig Information über Klassenzugehörigkeit verloren geht.

2.1 Klassifikationsverfahren

Die Aufgabe der Klassifikationsstufe ist die Zuordnung eines Musters $\rho_{\underline{f}}$, dargestellt durch einen Merkmalsvektor $\rho_{\underline{c}}$, zu einer von k bekannten Klassen Ω_k . Besondere Bedeutung kommt in der Mustererkennung der optimalen, statistischen Klassifikation zu, da diese so ausgelegt ist, daß das in einem konkreten Beispiel zu erwartende Risiko - also die Maßzahl für den Erfolg der Entscheidung - minimiert wird. Die optimale Klassifikation kann in der Anwendung normalerweise nicht erreicht werden, da die vorausgesetzten Kenntnisse über die Klassen nicht vorliegen. Der optimale Bayesklassifikator soll jedoch im folgenden dargestellt werden, da er zum einen die theoretisch bestmögliche Klassifikation beschreibt und zum anderen aus ihm über vereinfachende Annahmen andere Klassifikatoren abgeleitet werden können, die in der Anwendung durchaus eine Rolle spielen.

2.1.1 Der optimale Bayesklassifikator

Gegeben seien k Klassen $\Omega_1, \dots, \Omega_k$. Die klassenbedingten Dichten des Merkmalvektors $\underline{c}$

$$w(\underline{c} | \Omega_\kappa), \quad \kappa \in \{1, 2, \dots, k\}$$

seien bekannt, ebenso die a-priori Wahrscheinlichkeiten

$$p(\Omega_\kappa), \quad \kappa \in \{1, 2, \dots, k\}$$

die angeben, mit welcher Wahrscheinlichkeit ein Muster aus der Klasse Ω_κ beobachtet wird, das heißt, wie häufig Muster aus Klasse Ω_κ zu erwarten sind.

Sei nun

$$\delta(\Omega_\kappa | \underline{c}), \quad \kappa \in \{1, 2, \dots, k\}$$

eine Funktion, die für ein beliebiges $\underline{c} \in \mathbb{R}^n$ angibt, mit welcher Wahrscheinlichkeit ein Muster, das den Merkmalsvektor $\underline{c}$ liefert, zur Klasse Ω_κ gehört. Im folgenden wird eine solche Funktion eine Entscheidungsfunktion genannt. Gesucht wird eine Maßzahl dafür, wie gut eine Entscheidungsfunktion für ein konkretes Klassifikationsproblem geeignet ist. Im allgemeinen führt man in der statistischen Entscheidungstheorie das Risiko als eine solche Maßzahl ein.

Das Risiko, mit dem eine Entscheidungsfunktion - oder auch Entscheidungsregel genannt - behaftet ist, wird als Erwartungswert von Kosten definiert. Seien

$$r(\Omega_\kappa, \Omega_\lambda) \quad \kappa, \lambda \in \{1, 2, \dots, k\}$$

in einem speziellen Fall die Kosten, die entstehen würden, wenn ein Muster, das zu einer Klasse Ω_κ gehört, als zu Klasse Ω_λ gehörig erkannt wird. Für alle Klassifikationsprobleme sollte sinnvollerweise

$$r_{\min} \leq r(\Omega_\kappa, \Omega_\kappa) < r(\Omega_\kappa, \Omega_\lambda) \leq r_{\max} \quad \text{für } \kappa \neq \lambda \quad (2-1)$$

gelten, das heißt, die Kosten für eine richtige Entscheidung sollten niedriger sein als die Kosten für eine Fehlklassifikation.

Häufig wird eine Entscheidungsregel δ so konzipiert, daß Muster, die nicht mit genügender Sicherheit zu einer der k Klassen zugeordnet werden, "zurückgewiesen" werden. Formal wird dieser Vorgang durch die Einführung einer Rückweisungsklasse Ω_0 modelliert. Die Entscheidungsregel δ liefert dann für $\kappa = 0$ gerade die Wahrscheinlichkeit dafür, daß ein Muster nicht mit ausreichender Sicherheit in eine der Klassen Ω_1 bis Ω_k eingeordnet werden kann. Konzeptionell bereitet die Einführung einer Rückweisungsklasse keine Schwierigkeiten, im folgenden wird jedoch davon ausgegangen, daß eine Rückweisung ausgeschlossen ist. Man spricht in diesem Fall von erzwungener Entscheidung.

Mit den eingeführten Definitionen kann das Risiko $V(\delta)$ explizit als Funktion der Entscheidungsregel angegeben werden /N2, J1/:

$$V(\delta) = \int_{\mathbb{R}^n} \sum_{\lambda=1}^k \delta(\Omega_\lambda | \underline{c}) \cdot u_\lambda(\underline{c}) d\underline{c} \quad (2-2)$$

wobei

$$u_\lambda(\underline{c}) = \sum_{\kappa=1}^k r(\Omega_\kappa, \Omega_\lambda) \cdot w(\underline{c} | \Omega_\kappa) p(\Omega_\kappa) \quad (2-3)$$

Die Entscheidungsregel δ_0 , die $V(\delta)$ minimiert, ergibt sich zu /N2, J1/.

$$\delta_0(\Omega_\lambda | \underline{c}) = \begin{cases} 1 & \text{falls } u_\lambda(\underline{c}) = \min_{\kappa} u_\kappa(\underline{c}) \\ 0 & \text{sonst} \end{cases} \quad (2-4)$$

δ_0 ist also eine deterministische Entscheidungsregel. Aus (2-4) ist die prinzipielle Struktur der Klassifikationsstufe in Bild 1-1 ablesbar: Es sind die k Zahlen $u_\lambda(\underline{c})$ zu berechnen

und das zugehörige Muster der Klasse zuzuordnen, für die $u_k(\underline{c})$ minimal ist. Dieser Aufbau ist in Bild 2-1 schematisch dargestellt.

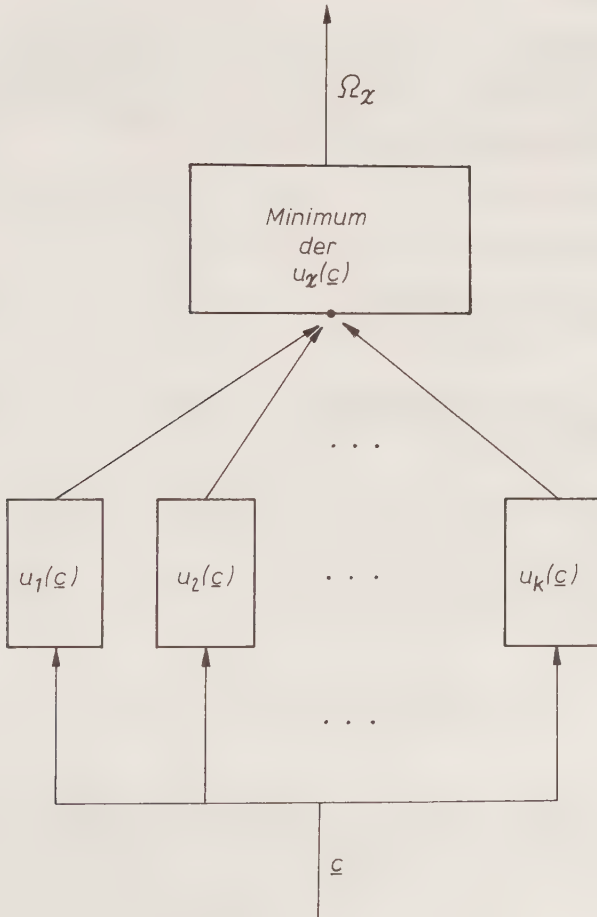


Bild 2-1 Struktur der Klassifikationsstufe

Die k Zahlen $u_k(\underline{c})$ sind die zentralen Größen für die Entscheidungsfindung. Da sie letztlich dazu berechnet werden, die Klassenzugehörigkeit festzustellen, nennt man sie Prüfgrößen /N2/.

Gleichung (2-4) zeigt eine weitere Wirkung der Verwendung der Bayesregel, wie die optimale Entscheidungsregel δ_0 auch genannt wird. Durch δ_0 wird eine Zerlegung des $\mathbb{R}^n$ in sogenannte Klassenbereiche induziert:

$$\Omega_\lambda^0 = \{\underline{c} | \delta_0(\Omega_\lambda | \underline{c}) = 1\} = \{\underline{c} | u_\lambda(\underline{c}) = \min_k u_k(\underline{c})\} \quad (2-5)$$

Damit schreibt sich das minimale Risiko als

$$V_{\min} = \sum_{\lambda=1}^k \int_{\Omega_\lambda^0} u_\lambda(\underline{c}) d\underline{c} \quad (2-6)$$

$V_{\min}$ heißt das Bayes-Risiko. Verwendet man die spezielle Kostenfunktion

$$\left. \begin{aligned} r(\Omega_\lambda, \Omega_\lambda) &= 0 \\ r(\Omega_\lambda, \Omega_\kappa) &= 1 \end{aligned} \right\} \quad \begin{aligned} \kappa, \lambda \in \{1, 2, \dots, k\} \\ \kappa \neq \lambda \end{aligned} \quad (2-7)$$

so wird aus (2-6) /N2, J1/:

$$V_{\min} = \sum_{\lambda=1}^k \sum_{\kappa \neq \lambda} p(\Omega_\kappa) \cdot \int_{\Omega_\lambda^0} w(\underline{c} | \Omega_\kappa) d\underline{c} \quad (2-8)$$

Das Integral in (2-8) ist die Wahrscheinlichkeit dafür, daß ein Muster aus Ω_κ im Klassenbereich Ω_λ^0 liegt. Dies ist die Wahrscheinlichkeit für eine Fehlklassifikation von Ω_κ nach Ω_λ . Das Risiko ist hier die Fehlerrate. Wegen der speziellen Wahl von δ_0 ist $V_{\min}$ die kleinste Fehlerrate, die in einem Klassifikationsproblem bei Zuhilfenahme statistischer Verfahren überhaupt erreichbar ist. Man nennt diese Fehlerrate die Bayesfehlerrate. In diesem Fall ist das Aufsuchen des

Minimums in (2-4) gleichwertig mit dem Aufsuchen der maximalen a posteriori Wahrscheinlichkeiten der Klassen $\Omega_k / N2$, J1/

$$\begin{aligned} u'_k(\underline{c}) &= p(\Omega_k | \underline{c}) \\ &= \frac{p(\Omega_k) \cdot w(\underline{c} | \Omega_k)}{w(\underline{c})} \end{aligned} \quad (2-9)$$

wobei

$$w(\underline{c}) = \sum_{\lambda=1}^k p(\Omega_\lambda) \cdot w(\underline{c} | \Omega_\lambda) \quad (2-10)$$

Wie bereits festgestellt wurde, liegt der Wert der Bayesregel zunächst im theoretischen Bereich. Dies liegt daran, daß im allgemeinen die Dichten $w(\underline{c} | \Omega_k)$ und die a priori Wahrscheinlichkeiten $p(\Omega_k)$ nicht bekannt sind. Die Aufgabe, die Dichten aus Stichproben zu schätzen, ist im allgemeinen sehr schwierig. Eine gängige Technik ist jedoch, vereinfachende Annahmen über die Dichten $w(\underline{c} | \Omega_k)$ zu machen. Es ist möglich, über eine Kette von vereinfachenden Annahmen über die Dichten $w(\underline{c} | \Omega_k)$ eine Hierarchie von Klassifikatoren abzuleiten, die im Anwendungsfall mit vertretbarem Aufwand an das konkrete Problem angepaßt werden können.

2.1.2 Einige Spezialisierungen des Bayesklassifikators

Eine naheliegende Vereinfachung ist, die allgemeinen Dichten $w(\underline{c} | \Omega_k)$ durch parametrische Formen zu ersetzen. Man macht also Annahmen über die Form der klassenbedingten Wahrscheinlichkeitsdichten und unterstellt, daß die Klassencharakteristika in den Parametern der Verteilung ausgedrückt werden können. Man gewinnt zwei Dinge: Zum ersten gibt es in der Statistik eine umfassende Theorie der Parameterschätzung aus Stichproben, zum anderen kann man hoffen, daß die Prüfgrößen $u'_k(\underline{c})$ in eine gut auswertbare Form überführt werden

können. Es soll darauf aufmerksam gemacht werden, daß ein durch solche Annahmen gewonnener Klassifikator für ein konkretes Problem genau dann ein Bayesklassifikator mit minimalem Risiko ist, wenn die betrachteten Klassen statistische Eigenschaften besitzen, die der vereinfachenden Annahme entsprechen. Im Anwendungsfall sind solche Annahmen über statistische Eigenschaften von Zufallsvariablen im Einzelfall natürlich nachzuprüfen, z.B. mit Hilfe statistischer Tests und einer Stichprobe. In der vorliegenden Arbeit wird jedoch auch der Standpunkt vertreten, daß durch Annahmen über die Form von klassenbedingten Dichten Klassifikatoren aus dem Bayesklassifikator abgeleitet werden können, die per se als Klassifikatoren bzw. Klassifikatorstrukturen angesehen werden. Solche Klassifikatoren werden in aller Regel durch die zu berechnenden Prüfgrößen definiert. Man kann auf solche Art erhaltene Klassifikatoren in einem Anwendungsfall einsetzen, ohne genauere Einsicht in die Art der klassenbedingten Dichten zu besitzen. Die Leistung eines solchen Klassifikators ist dann eben nicht optimal im Bayes'schen Sinn. Dafür mag im konkreten Fall seine Anwendung gegenüber dem problemgerechten Bayesklassifikator erheblich einfacher sein. Die im folgenden vorgestellten Vereinfachungen liefern in diesem Sinn also eine Palette möglicher Klassifikatoren.

2.1.2.1 Normalverteilungsklassifikator

Nimmt man an, daß alle klassenbedingten Dichten normale Dichten mit den Mittelvektoren $\underline{\mu}_k$ und den Kovarianzmatrizen $\underline{K}_k$ sind, daß also

$$w(\underline{c} | \Omega_k) = \frac{1}{\sqrt{|2\pi \underline{K}_k|}} \exp\left\{-\frac{1}{2}(\underline{c} - \underline{\mu}_k)' \underline{K}_k^{-1} (\underline{c} - \underline{\mu}_k)\right\} \quad (2-11)$$

gilt, so kann man aus (2-9) die Prüfgrößen $u_k(\underline{c})$ für den

Normalverteilungsklassifikator (im folgenden mit NVK abgekürzt) ableiten:

$$u_k(\underline{c}) = \frac{p(\Omega_k)}{\sqrt{|2\pi K_k|}} \exp\left\{-\frac{1}{2}(\underline{c} - \underline{\mu}_k)' t_{K_k}^{-1} (\underline{c} - \underline{\mu}_k)\right\} \quad (2-12)$$

Üblicherweise berechnet man den Logarithmus dieser Prüfgrößen. Man erhält dann

$$u'_k(\underline{c}) = \ln u_k(\underline{c}) = -\frac{1}{2}(\underline{c} - \underline{\mu}_k)' t_{K_k}^{-1} (\underline{c} - \underline{\mu}_k) + \gamma_k \quad (2-13)$$

mit

$$\gamma_k = \ln p(\Omega_k) - \frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |K_k|$$

Es ist jetzt möglich, eine genauere Aussage über die Menge aller $\underline{c}$ zu machen, für die

$$u'_k(\underline{c}) = u'_\lambda(\underline{c}) \quad (2-14)$$

gilt. Aus (2-13) folgt unmittelbar, daß diese $\underline{c}$ alle auf einem Kegelschnitt liegen, der durch die Gleichung (2-14) definiert ist. Dieser Kegelschnitt definiert - geometrisch veranschaulicht - die Klassengrenze zwischen den Klassen Ω_k und Ω_λ . Ein Kegelschnitt im $\mathbb{R}^n$ ist eine $(n-1)$ -dimensionale Hyperfläche. Die Wahrscheinlichkeit, daß ein Muster auftritt, dessen Merkmalsvektor auf dieser Fläche liegt, ist somit - Normalverteilungen vorausgesetzt - Null.

2.1.2.2 Abstandmessende Klassifikatoren

In Gleichung (2-13) tritt der Teilausdruck

$$(\underline{c} - \underline{\mu}_k)' t_{K_k}^{-1} (\underline{c} - \underline{\mu}_k) \quad (2-15)$$

auf. Da K_k als Kovarianzmatrix symmetrisch und reellwertig

ist, ist $\underline{K}_\kappa$ und damit auch $\underline{K}_\kappa^{-1}$ positiv definiert. (2-15) kann deshalb als ein Abstandsmaß aufgefaßt werden. (2-15) gibt das Quadrat des Abstandes von $\underline{c}$ zu $\underline{\mu}_\kappa$ im $\mathbb{R}^n$ an, wobei eine Metrik verwendet wird, die eine Gewichtung beinhaltet. Es ist nun möglich, weitere Annahmen über die Art der klassenbedingten Dichten der Merkmalvektoren zu machen, die zur Folge haben, daß die Entscheidung über Klassenzugehörigkeit auf dem Weg über Abstandsmessungen zu Klassenrepräsentanten getroffen werden kann. Die statistischen Klassifikatoren erhalten auf diesem Weg eine geometrische Veranschaulichung im $\mathbb{R}^n$.

Einen ersten abstandsmessenden Klassifikator erhält man, wenn man die Zahlen γ_κ aus Gleichung (2-13) vernachlässigt. Man erhält dann die Prüfgrößen

$$u_\kappa(\underline{c}) = (\underline{c} - \underline{\mu}_\kappa) \underline{K}_\kappa^{-1} (\underline{c} - \underline{\mu}_\kappa) \quad (2-16)$$

Ausgewählt wird die Klasse mit minimalem $u_\kappa(\underline{c})$. Der resultierende Klassifikator wird mit "modifizierter Minimum-Abstand-Klassifikator" bezeichnet (abgekürzt MMA). Bei diesem Klassifikator wird jeder der Klassen Ω_κ eine besondere Metrik zugeordnet und mit Hilfe dieser Metrik der (klassenspezifische) Abstand zum Klassenschwerpunkt gemessen. Die Klassengrenzen bleiben Flächen zweiter Ordnung.

Nimmt man nun an, daß alle klassenspezifischen Kovarianzmatrizen übereinstimmen, so erhält man einen Klassifikator, der als Klassengrenzen nur noch Flächen erster Ordnung, also Hyperebenen, kennt. Ist nämlich für alle Klassen $\underline{K}_\kappa = \underline{K}$, so wird aus (2-16) beziehungsweise aus (2-13) unter der zusätzlichen Annahme gleicher a priori Wahrscheinlichkeiten $p(\Omega_\kappa)$

$$u_\kappa(\underline{c}) = (\underline{c} - \underline{\mu}_\kappa) \underline{K}^{-1} (\underline{c} - \underline{\mu}_\kappa) \quad (2-17)$$

und hieraus

$$u_{\kappa}(\underline{c}) = \underline{c}_t \underline{K}^{-1} \underline{c} - 2 \underline{c}_t \underline{K}^{-1} \underline{\mu}_{\kappa} + \underline{\mu}_{\kappa} \underline{K}^{-1} \underline{\mu}_{\kappa}$$

Die Klassengrenze erfüllt also die Gleichung

$$\underline{\mu}_{\kappa} \underline{K}^{-1} \underline{\mu}_{\kappa} - 2 \underline{c}_t \underline{K}^{-1} \underline{\mu}_{\kappa} = \underline{\mu}_{\lambda} \underline{K}^{-1} \underline{\mu}_{\lambda} - 2 \underline{c}_t \underline{K}^{-1} \underline{\mu}_{\lambda} ,$$

eine lineare Gleichung in $\underline{c}$. Der so definierte Klassifikator wird als Mahalanobis-Abstand-Klassifikator (MHK) bezeichnet.

Ist zusätzlich noch $\underline{K}$ die Einheitsmatrix, so wird aus dem MHK der einfache Minimum-Abstand-Klassifikator (MAK), ein Klassifikator, der im wesentlichen einen Vergleich zwischen einem Muster und k Klassenrepräsentanten realisiert. Dies ist sicherlich eine einfache und einleuchtende Prozedur. Die Zusammenhänge zwischen den bislang vorgestellten Klassifikatoren sind summarisch im Bild 2-2 dargestellt.

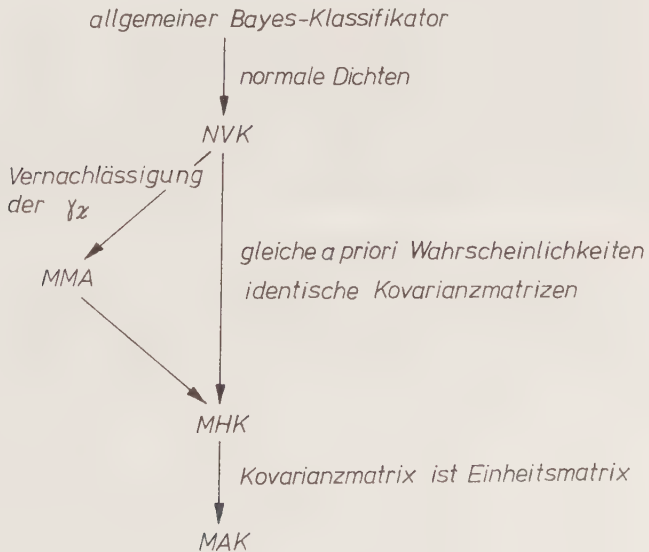


Bild 2-2 Spezialisierungen des Bayesklassifikators

2.1.3 Der im quadratischen Mittel angepaßte Klassifikator

Ein Nachteil des entscheidungstheoretischen Ansatzes, der zur Bestimmung des Bayesklassifikators und seiner Spezialfälle führte, ist, daß explizite Voraussetzungen über die klassenbedingten Dichten $w(\underline{c}|\Omega_k)$ gemacht werden mußten bzw. daß diese Dichten als bekannt vorausgesetzt wurden. Die Methode, einen Klassifikator "im quadratischen Mittel" anzupassen, verzichtet ganz auf die Einbeziehung von Dichten und versucht statt dessen, direkt einen Bezug zwischen Klassifikationsleistung und den verwendeten Prüfgrößen herzustellen. Der Ansatz führt auf eine Optimierungsaufgabe. Der prinzipielle Weg soll im folgenden kurz dargestellt werden, eine ausführliche Darstellung findet sich in /S1/.

Gegeben seien wieder k Klassen Ω_k . In Anlehnung an die Struktur des Klassifikators aus Bild 2-1 wird angestrebt, k Trennfunktionen $d_k(\underline{c})$ zu bestimmen, die im Idealfall die Eigenschaft haben sollen, daß $d_\lambda(\underline{c})$ genau dann den Wert 1 annimmt, wenn $\underline{c}$ zu einem Muster aus Klasse Ω_λ gehört. Für alle anderen $\underline{c}$ wird gewünscht, daß $d_\lambda(\underline{c})$ Null wird. Dieses Ziel ist in der Praxis in aller Regel nicht realisierbar, eine Annäherung kann aber erreicht werden, wenn man die zu diesem Ziel gehörige Optimierungsaufgabe mit der Zielfunktion

$$S^2 = E\{|\underline{d}(\underline{c}) - \underline{y}|^2\} \stackrel{!}{=} \text{Min} \quad (2-18)$$

löst. In (2-18) ist $\underline{d}(\underline{c})$ der k -dimensionale Vektor

$$\underline{d}(\underline{c}) = \begin{pmatrix} d_1(\underline{c}) \\ d_2(\underline{c}) \\ \vdots \\ d_k(\underline{c}) \end{pmatrix} \quad (2-19)$$

und $\underline{y}$ ein k -dimensionaler Einheitsvektor, der eine Eins in

genau der Komponente trägt, die der richtigen Klasse von $\underline{c}$ entspricht. Der Erwartungswert ist bezüglich der Verteilung der Zufallsvariablen "Merkmalsvektor $\underline{c}$ " zu bilden.

Es kann nun über die Art der Trennfunktionen d_k verfügt werden. Eine einfache und gängige Struktur des Klassifikators ergibt sich, wenn man $d_k(\underline{c})$ als Linearkombination der Komponenten des Merkmalvektors festlegt

$$d_k(\underline{c}) = \underline{a}_{k,t} \cdot \underline{c} \quad (2-20)$$

Definiert man

$$\underline{A} = (\underline{a}_1, \underline{a}_2, \dots, \underline{a}_k) \quad (2-21)$$

so schreibt sich (2-18) als

$$S^2 = E\{|\underline{A}_t \underline{c} - \underline{y}|^2\} \stackrel{!}{=} \text{Min}$$

wo die Minimierungsaufgabe nun in Abhängigkeit von $\underline{A}$ zu lösen ist. Die Lösung kann explizit angegeben werden /S1/:

$$\underline{A} = E\{\underline{c}\underline{c}_t\}^{-1} \cdot E\{\underline{c}\underline{y}_t\} \quad (2-22)$$

Es bleiben noch die Erwartungswerte zu berechnen. Dies gelingt mit Hilfe einer klassifizierten Stichprobe, wenn man berücksichtigt, daß $E\{\underline{c}\underline{c}_t\}$ die Kovarianzmatrix der Verteilung des Merkmalsvektors $\underline{c}$ ungeachtet jeder Klassenzugehörigkeit ist und

$$E\{\underline{c}\underline{y}_t\} = \left(p(\Omega_1) \underline{y}_1, \dots, p(\Omega_k) \underline{y}_k \right)$$

ist, also im wesentlichen die Matrix der klassenbedingten Mittelvektoren, jeder gewichtet mit der a priori Wahrscheinlichkeit der Klassen.

Zunächst scheint der resultierende Quadrat-Mittel-Klassifikator (QMK) von wesentlich anderem Typ zu sein als der optimale Bayesklassifikator. Es kann aber gezeigt werden /S1/, daß die $d_k(\underline{c})$ als Schätzungen für die $p(\Omega_k | \underline{c})$ interpretiert werden dürfen. Von der Theorie her wird also in beiden Fällen das gleiche Ziel erreicht, in der Praxis hat man es jedoch mit unterschiedlichen Verfahren einmal bei der Klassifikation eines Musters und zum anderen bei der Anpassung des Klassifikationsverfahrens zu tun. Letztendlich muß natürlich berücksichtigt werden, daß ein Klassifikator nach (2-20) im Merkmalsraum nur Trennflächen erster Ordnung realisieren kann. Dies ist jedoch kein prinzipieller Nachteil, da durch Ersetzen von $\underline{c}$ durch ein $\tilde{\underline{c}}$, das als Komponenten auch Produkte von Merkmalen enthalten darf, implizit auch Trennflächen zweiter oder höherer Ordnung erreicht werden können.

2.1.4 Hinweise auf weitere Klassifikatortypen

Die bisher beschriebenen Klassifikatoren spielen in der vorliegenden Arbeit eine besondere Rolle, da sie in den Experimenten zum Einsatz kommen. Außer diesen gibt es noch eine Reihe anderer Klassifikationsschemata. Übersichten finden sich zum Beispiel in /H1, K1, N2, D1/. Als wesentlichste Vertreter seien hier die geometrischen Klassifikatoren genannt /N2/, zu denen auch der Quadrat-Mittel-Klassifikator gerechnet werden muß, weiterhin die Nächste-Nachbar-Regel /C1/, die sequentiellen Klassifikatoren /F2/, die hierarchischen Klassifikatoren /K3/ und als nichtnumerische Verfahren die "syntaktischen" Klassifikatoren /F3/.

2.2 Merkmalsbewertung und Merkmalsauswahl

Während man die Theorie der numerischen Klassifikation

inzwischen als relativ stabil bezeichnen kann, wie auch die inzwischen große Zahl von Lehrbüchern andeutet /F1, N2, D1, Y1, T2, P1/, unterliegt die Theorie der Merkmalsauswahl noch immer ständigen Änderungen. Man hat zwar immer wieder versucht, durch Übersichtsartikel wie /K1, V1, C2, K2/ Ordnung in die Theorie zu bringen, aber bis heute ist nicht abzusehen, daß befriedigende Lösungen gefunden werden können. Dies besagt nicht, daß sich trotzdem Standardmethoden wie die Karhunen-Loève-Transformation /F4/ etabliert haben.

Wie bereits festgestellt wurde, hat man die Theorie der Merkmalsauswahl getrennt von der Theorie der Klassifikation vorangetrieben. Damit stellt sich natürlich zunächst die Frage, wie man einzelne Merkmale oder einen Merkmalvektor hinsichtlich der Eignung zur Unterscheidung von Klassen beurteilen kann. Ein großer Teil der Literatur zur Merkmalsauswahl beschäftigt sich genau mit diesem Thema, wobei in die Betrachtung mit eingeschlossen wird, wie aus einer Maßzahl für die Güte von Merkmalen eventuell auf die Bayes-Fehlerrate des Problems geschlossen werden kann /C2, T1, K1/.

Neben diesen Gütemaßen wird eine Auswahlprozedur benötigt, die angibt, in welcher Art und Weise aus den vorhandenen n Merkmalen n' Merkmale erzeugt werden sollen. Diese beiden Teilaspekte werden in jeweils einem eigenen Abschnitt dargestellt.

2.2.1 Bewertungsmaße für Merkmale

Im folgenden werden Bewertungsmaße angegeben, die ein Zweiklassen-Problem voraussetzen. Eine Verallgemeinerung für k Klassen erhält man in allen Fällen durch Mittelung der Maßzahlen über alle $\binom{k}{2}$ Klassenpaare. Die aus der Literatur bekannten Bewertungsmaße kann man gemäß den ihnen zugrunde

liegenden heuristischen Überlegungen in fünf große Gruppen gliedern:

- a) Gütekriterien, die auf der Messung von Punktabständen in einer klassifizierten Stichprobe aufbauen /N2/.
- b) Gütekriterien, die auf Quasimetriken für Wahrscheinlichkeitsdichten aufbauen /K2/.
- c) Gütekriterien, die den Grad der Unabhängigkeit des Merkmalvektors $\underline{c}$ von der Klassenzugehörigkeit messen /V1/.
- d) Gütekriterien, die auf dem von einer Stichprobe belegten Volumen im n-dimensionalen Raum aufbauen /P1/.
- e) Gütekriterien, die informationstheoretische Maße verwenden /C2/.

Es würde zu weit führen, die einzelnen Gruppen mit den zwischen ihnen bestehenden Zusammenhängen vorzustellen. Die Gruppen seien deshalb jeweils nur exemplarisch dargestellt.

Zu a)

Es ist intuitiv einleuchtend, daß die Klassifikation von Mustern leicht wird, wenn im Merkmalraum die Punkte einer Klasse jeweils ein Häufungsgebiet bilden und die Häufungsgebiete verschiedener Klassen weit auseinander liegen. Ein Gütemaß, das diese Vorstellung in eine Maßzahl umsetzt, ist

$$G = \frac{1}{2N_K N_\lambda} \sum_{k=1}^{N_K} \sum_{m=1}^{N_\lambda} d(\underline{c}_k, \underline{c}_\lambda)$$

wo $d(.,.)$ eine geeignete Metrik im $\mathbb{R}^n$ ist. G ist der mittlere Abstand zwischen zwei Mustern aus verschiedenen Klassen.

zu b)

Eine naheliegende Idee ist, daß ein Klassifikationsproblem dann leicht zu lösen sein müßte, wenn die klassenbedingten Dichten des Merkmalvektors in irgendeinem Sinne sehr unterschiedlich sind. Der Unterschied soll mit Hilfe eines Abstandes im Raum aller Wahrscheinlichkeitsdichten ausgedrückt

werden. Man kennt zwar keine Maßzahl, die aus zwei Dichten berechnet werden kann und alle Eigenschaften einer Metrik besitzt, jedoch kann man eine ganze Reihe von Quasimetriken angeben, die allerdings alle die Dreiecksungleichung nicht erfüllen. Tabelle 2-1 gibt einen Eindruck von der Vielfalt der möglichen Quasimetriken.

Kolmogoroff-Abstand	$\int p_1(\underline{c}) - p_2(\underline{c}) d\underline{c}$
Bhattacharryya-Abstand	$-\ln \int \sqrt{p_1(\underline{c}) p_2(\underline{c})} d\underline{c}$
Matusita-Abstand	$\sqrt{\int (\sqrt{p_1(\underline{c})} - \sqrt{p_2(\underline{c})})^2 d\underline{c}}$
Abstandsmaß von Lissack und Fu	$\int p_1(\underline{c}) - p_2(\underline{c}) ^\alpha d\underline{c}$
Chernoff-Abstand	$-\ln \int p_1^\alpha(\underline{c}) p_2^{1-\alpha}(\underline{c}) d\underline{c}$
Patrick-Fisher-Abstand	$\sqrt{\int (p_1(\underline{c}) - p_2(\underline{c}))^2 d\underline{c}}$
Divergenz	$\int (p_1(\underline{c}) - p_2(\underline{c})) \ln \frac{p_1(\underline{c})}{p_2(\underline{c})} d\underline{c}$
quadratischer Abstand	$\int (p_1(\underline{c}) - p_2(\underline{c}))^2 d\underline{c}$

Tabelle 2-1

Die notwendigen mehrdimensionalen Integrationen können bei bekannten Dichten p_1 und p_2 unter Umständen analytisch ausgeführt werden. Die Maßzahlen können in solchen Fällen günstig berechnet werden.

zu c)

Die Grundidee bei diesem Verfahren ist, die Verbundwahrscheinlichkeit von Merkmalvektor und Klassenzugehörigkeit mit dem Produkt der entsprechenden Randverteilungen zu vergleichen. Stimmen die verglichenen Verteilungen überein, so sind Merkmalvektor und Klassenzugehörigkeit statistisch unabhängig;

das bedeutet, daß aus der Kenntnis des Merkmalvektors nicht auf die Klassenzugehörigkeit geschlossen werden kann. Maßzahlen können über die Abstandsmaße aus Tabelle 2-1 gewonnen werden.

zu d)

Die Grundidee ist wie die der Verfahren zu a), nur daß als Maßzahl die sogenannte Scatterzahl herangezogen wird. Die Scatterzahl einer Punktmenge im n-dimensionalen Raum, bezogen auf einen Pivotpunkt ist die Summe der Volumina aller n-dimensionalen Parallelotope aus n Punkten der Punktmenge und dem Pivotpunkt. Wählt man als Pivotpunkt den Schwerpunkt der Punktmenge, erhält man als Scatterzahl die Determinante der empirischen Kovarianzmatrix.

Zu e)

Die Grundidee informationstheoretischer Maße ist, die Information zu messen, die die Beobachtung des Merkmalvektors über die Klassenzugehörigkeit liefert. Typische Maße dieser Art sind die Transinformation

$$G = \sum_{\kappa=1}^k \int_{\mathbb{R}^n} p(\Omega_{\kappa}) \cdot w(\underline{c} | \Omega_{\kappa}) \log \frac{w(\underline{c} | \Omega_{\kappa})}{w(\underline{c})} d\underline{c}$$

oder die Entropie

$$G = - \sum_{\kappa=1}^k \int_{\mathbb{R}^n} P(\Omega_{\kappa} | \underline{c}) w(\underline{c}) \log P(\Omega_{\kappa} | \underline{c}) d\underline{c}$$

Die kurz umrissenen Gütekriterien gestatten, einen Satz von Merkmalen hinsichtlich seiner Eignung für die Klassifikation in einem intuitiv einleuchtendem Sinn zu bewerten. Prinzipiell mißt man den Wert eines solchen Kriteriums für die Merkmalsauswahl an folgenden Forderungen:

- i) die Berechnung des Gütemaßes sollte mit geringem Aufwand möglich sein, da die Durchführung der Merkmalsauswahl in aller Regel die Bewertung sehr vieler Merkmalsätze erfordert
- ii) das Gütemaß sollte, wenn schon kein direkter Bezug zum Risiko hergestellt werden kann, wenigstens eine anschauliche Bedeutung haben
- iii) das Gütemaß sollte in Relation zum Risiko stehen; zum Beispiel wäre ein monotoner, funktioneller Zusammenhang mit dem Klassifikationsfehler nützlich
- iv) in den Werten des Gütemaßes sollte sich ausdrücken, inwieweit Klassenbereiche überlappen
- v) quasimetrische Eigenschaften des Gütemaßes könnten die Anwendung der Theorie der Optimierung in metrischen Räumen nahelegen
- vi) die Bewertung sollte die Identifikation der Beiträge eines Merkmals zur Bewertung eines ganzen Satzes von Merkmalen zulassen. In diesem Fall bietet sich als Auswahlverfahren zum Beispiel dynamische Programmierung an
- vii) der Wert des Gütemaßes sollte sich nicht ändern, wenn man den Merkmalsraum einer bijektiven Abbildung unterwirft.

Die Eigenschaften (i) bis (vii) sind nicht zwingend, helfen aber bei der Auswahl eines speziellen Gütekriteriums /F1/. Es bleibt nun darzulegen, wie diese Gütekriterien bei der Merkmalsauswahl eingesetzt werden.

2.2.2 Verfahren zur Merkmalsauswahl

Liegt eine große Anzahl n von meßbaren Eigenschaften von Mustern vor, so gibt es zwei prinzipiell verschiedene Möglichkeiten, n' Merkmale für die Klassifikation zu erzeugen. Zum einen kann man nach Untermengen des großen Merkmalsatzes suchen, die ein vorgegebenes Gütekriterium, etwa von der Art wie in 2.2.1 vorgestellt, optimieren, zum

anderen kann man aber auch nach einer vektorwertigen Funktion von n Variablen suchen und eventuelle Parameter einer solchen Funktion so einstellen, daß wiederum ein Gütekriterium optimal ausfällt.

2.2.2.1 Auswahlverfahren

Auswahlverfahren bestimmen eine optimale Untermenge mit n' Merkmalen, wenn n Merkmale überhaupt als Kandidaten zur Verfügung stehen. Im allgemeinen verbietet sich die Untersuchung aller $\binom{n}{n'}$ Untermengen aus Aufwandsgründen, so daß hier suboptimale Suchstrategien eingesetzt werden müssen. Verwendet man Gütemaße mit der Eigenschaft (vi) aus Kapitel 2.2.1, so können mit Erfolg Methoden der dynamischen Programmierung eingesetzt werden /C3/. In anderen Fällen behilft man sich mit der Bewertung einzelner Merkmale oder sucht aufsteigende Mengen von Merkmalen /J1/. Man muß sich jedoch mit der Erkenntnis abfinden, daß der Satz aus den zwei am besten bewerteten Merkmalen nicht der beste Satz aus zwei Merkmalen sein muß, auch dann nicht, wenn man Annahmen wie statistische Unabhängigkeit von Merkmalen hinzunimmt /C4, E1/.

2.2.2.2 Merkmalsauswahl durch Transformationen

Merkmalsauswahl durch Transformation versucht i. a. durch Lösen einer Optimierungsaufgabe aus einer parametrischen Schar von Abbildungen des $\mathbb{R}^n$ in den $\mathbb{R}^{n'}$ einen n' -dimensionalen Merkmalvektor zu finden, der möglichst gut für die Klassifikation geeignet ist. Man kann Auswahlverfahren nach Abschnitt 2.2.2.1 als Projektionen auffassen und als solche in die hier vorgestellten Transformationen mit einreihen; man hätte jedoch in diesem Fall unter Umständen Restriktionen bei der Lösung der Optimierungsaufgabe zu berücksichtigen, die die spezielle Form dieser Auswahlmatrizen sicher-

stellen. Es erweisen sich dann eben andere Verfahren als zweckmäßiger.

Im Rahmen der vorliegenden Arbeit werden lineare Transformationen der Art

$$\underline{c}' = \underline{A}_t \cdot \underline{c} \quad (2-23)$$

betrachtet, wo $\underline{A}$ eine $n \times n'$ -Matrix ist. Gesucht wird bei dieser sogenannten "linearen Merkmalsauswahl" die Matrix $\underline{A}$, für die ein Gütekriterium im Zielraum optimal wird. Die Gütekriterien aus Abschnitt 2.2.1 wären also als Funktionen von $\underline{A}$ auszudrücken und dann - bei bekannten statistischen Parametern und geeigneten Einschränkungen über die Art der Matrix $\underline{A}$ - über die Menge aller zulässigen $\underline{A}$ zu optimieren.

Zu diesem Thema gibt es eine Reihe von Resultaten. Viele sind analytischer Natur, d. h. man kennt geschlossene Ausdrücke zur Berechnung der optimalen Transformationsmatrix $\underline{A} /F1/$. Fast keine Ergebnisse sind für die Fälle bekannt, wo $\underline{A}$ mit einem iterativen Verfahren berechnet werden muß. In $/F1/$ und $/T2/$ werden zum Beispiel eine Reihe von Gütekriterien ausgewertet, wo die Transformationsmatrix $\underline{A}$ als Lösung einer Eigenwertaufgabe formuliert werden kann. Allerdings geschieht dies häufig unter der zusätzlichen Annahme, daß ein Zwei-Klassen-Problem vorliegt. Numerische Methoden werden in $/D2/$ am Beispiel Divergenz vorgestellt.

Eine besonders einfache Methode, die Transformationsmatrix $\underline{A}$ festzulegen, ist, den Merkmalsvektor $\underline{c}$ nach einer orthormalen Basis des $\mathbb{R}^n$ zu entwickeln und nur einen Teil der Entwicklungskoeffizienten als neue Merkmale zu verwenden. Die Matrix $\underline{A}$ hat dann als Spalten die Vektoren der Orthormalbasis. Man kann dann verschiedenste, aus der Mathematik bekannte Reihenentwicklungen zur Merkmalsauswahl heranziehen. Grundsätzlich kann man zwei Arten von Reihenentwicklungen unterscheiden $/N2/$: problemabhängige und problem-

unabhängige Reihenentwicklung. Im wesentlichen läuft der Unterschied darauf hinaus, daß bei problemabhängigen Reihenentwicklungen eine Optimierungsaufgabe angegeben werden kann, deren Lösung die angestrebte, problemabhängige Reihenentwicklung ist und in deren Zielfunktion Kenngrößen des zu behandelnden Problems wie empirische Kovarianzmatrizen eingehen. Die problemunabhängigen Reihenentwicklungen dagegen, wie zum Beispiel die Fourierreihenentwicklung, werden nach rein mathematischen Kriterien aufgebaut. Beide Methoden haben je ihre eigenen Vorteile für die Mustererkennung: während problemabhängige Reihenentwicklungen für ein vorgegebenes Problem optimal gestaltet werden können, kennt man bei problemunabhängigen Reihenentwicklungen häufig schnelle Algorithmen /A1/, eine Eigenschaft, die im Lichte großer Datenmengen bei hohem Durchsatz nicht uninteressant ist.

Eine spezielle problemorientierte Reihenentwicklung hat sich in der Mustererkennung zu einem Standardverfahren entwickelt. Da sie auch im Rahmen der vorliegenden Arbeit eine besondere Rolle spielt, soll diese Transformation etwas detaillierter dargestellt werden.

2.2.2.3 Karhunen-Loève-Transformation

Die Karhunen-Loève-Transformation (KL-Transformation) ist bekannt geworden, weil sie eine Reihe von Optimalitätseigenschaften aufweist. Es ist zum einen gezeigt worden, daß bei Vorgabe dieser Optimalitätskriterien auf analytischem Wege das Karhunen-Loève-Koordinatensystem folgt /N2/. Es wurde aber auch bewiesen, daß das Karhunen-Loève-Koordinatensystem eindeutig ist und die unten noch anzugebenden Eigenschaften aufweist /W1/.

Das Karhunen-Loève-Koordinatensystem einer vorgegebenen Stichprobe besteht aus den Eigenvektoren der empirischen

Kovarianzmatrix der Stichprobe (die Klassenzugehörigkeit der Stichprobenelemente geht in diesem Fall nicht mit ein). Bekanntlich hat eine Kovarianzmatrix als reelle symmetrische Matrix immer einen vollständigen Satz von orthonormierten Eigenvektoren /Z1/. Die Eigenvektoren der empirischen Kovarianzmatrix werden so geordnet, daß die zugehörigen Eigenwerte eine monoton fallende Zahlenfolge bilden. Unter dieser Voraussetzung besitzt die KL-Entwicklung folgende Eigenschaften:

- a) Sie minimiert den mittleren Approximationsfehler beim Abbruch der Entwicklung nach $n' < n$ Gliedern. Als Approximationsfehler wird das Betragsquadrat des Fehlers nach der Rekonstruktion verwendet, gemittelt über alle Stichprobenelemente.
- b) Die Entropiefunktion

$$-\sum_{i=1}^n \sigma_i^2 \ln \sigma_i^2$$

zu den Varianzen aller n Entwicklungskoeffizienten c_i' wird minimal, d. h. die KL-Transformation konzentriert die "Gesamtvarianz" einer Stichprobe auf möglichst wenig Koordinaten.

- c) Die mit Hilfe der KL-Transformation erzeugten Merkmale sind unkorreliert.

Weitere Eigenschaften der KL-Transformation finden sich in der Literatur /W1, F4, N2/. Die KL-Transformation wird gerne als Grundlage für weitere Verfahren zur Dimensionsreduktion verwendet, bekannt geworden sind z. B. SELFIC und CLAFIC /W2/, dient aber auch einfach als Vergleichsverfahren /A1/. Im Rahmen dieser Arbeit kommen ihr beide Aufgaben zu: Zum einen wird sie benutzt, um vorab die Dimension von Merkmalsvektoren zu reduzieren, zum anderen, um einen Vergleich eines Standardverfahrens mit den hier entwickelten durchführen zu können.

3. Klassifikatorbezogene Merkmalsauswahl

3.1 Das Konzept

In Kapitel 1 wurde auf den prinzipiellen Aufbau eines Klassifikationssystems eingegangen. Es wurde festgestellt, daß solch ein System eine Merkmalsauswahlstufe und eine Klassifikationsstufe beinhaltet. Aufgabe der Klassifikationsstufe ist die Zuordnung eines Musters zu einer von k möglichen Klassen aufgrund der Kenntnis von Merkmalvektor und Klassencharakteristika. Aufgabe der Merkmalsauswahlstufe ist zum einen die Bestimmung von Merkmalen an einem gegebenen Muster, zum anderen die Reduktion einer großen Zahl von Merkmalen auf einige wenige, die jedoch noch soviel Information über die Klassenzugehörigkeit des Musters enthalten sollen, wie irgend möglich ist.

Der Entwurf eines solchen Systems zerfällt grundsätzlich in zwei Teile, die bislang völlig oder fast völlig getrennt voneinander behandelt werden: den Entwurf der Merkmalsextraktionsstufe und die Anpassung der Klassifikationsstufe. In Kapitel 2 wurden einige der bekannten Verfahren summarisch dargestellt. Ein Nachteil dieser Vorgehensweise ist, daß Aussagen über die Leistungsfähigkeit des Gesamtsystems aus einer speziellen Merkmalsauswahlstufe und einer Klassifikationsstufe so gut wie nicht gewonnen werden können. Natürlich kann man aus den Werten der Abstandsmaße aus Kapitel 2.2.2 Schranken für Fehlerraten bei Verwendung eines festgelegten Satzes von Merkmalen gewinnen /K1/, man darf aber dabei nicht übersehen, daß

- a) die Abstandsmaße nur dann berechnet werden können, wenn die entsprechenden Dichten bekannt sind,
- b) die Abstandsmaße für allgemeine Dichten schwer berechnet werden können,
- c) die Abstandsmaße bei einschränkenden Annahmen über die

Form der Dichten zwar leichter berechnet werden können (z. B. sind der Bhattacharyya-Abstand und die Divergenz für normale Dichten als Ausdrücke darstellbar, in denen keine Integration mehr vorkommt), daß aber dann auch das zu bearbeitende Klassifikationsproblem den vereinfachenden Annahmen genügen muß,

- d) die Fehlerschranken nur für den Bayesfehler gelten.

Im Normalfall steht der Konstrukteur eines Klassifikationssystems vor dem Problem, mit relativ wenigen und unpräzisen Kenntnissen über den Aufbau der Muster und der Klassen ein funktionstüchtiges System bauen zu müssen. So zum Beispiel besteht die Kenntnis über Muster und Klassen häufig aus einer klassifizierten Stichprobe, nämlich einer Menge von Mustern, die ein "unfehlbarer Lehrer" mit den richtigen Klassennamen gekennzeichnet hat. Diese Stichprobe muß für den Entwurf und die Feststellung der Leistung des Systems ausreichen. Eine beliebte Vorgehensweise in solch einer Situation, in der über Dichten, a priori Wahrscheinlichkeiten, Kosten für Fehlklassifikation und ähnliches nur vage Vorstellungen existieren, ist, ein bewährtes Klassifikationschema, etwa nach Kapitel 2.1, fest vorzugeben und diesem Klassifikator Merkmale anzubieten, die sich aus den Koeffizienten irgendeiner Reihenentwicklung ergeben. Die Struktur und die Arbeitsweise des Klassifikators werden nach dessen Vorgabe nicht weiter betrachtet.

Es wird im Rahmen dieser Arbeit vorgeschlagen, Merkmalsauswahlstufe und Klassifikationsstufe nicht mehr getrennt voneinander zu sehen, sondern quasi eine Brücke zwischen den beiden Teilbereichen aufzubauen. Konkret könnte eine solche Brücke zwei verschiedene Gestalten annehmen:

- a) man könnte versuchen, bei vorgegebenem Merkmalsvektor ein Klassifikationsschema zu finden, das die Leistung des Gesamtsystems optimiert
- b) bei vorgegebener Klassifikatorstruktur könnte man versuchen, Merkmale zu finden, die für diesen speziellen Klassifikator besonders gut geeignet sind.

Die praktische Berechtigung beider Verfahren kann z. B. in den Kosten für die Realisierung des Gesamtsystems liegen. Es sind sowohl einfach zu realisierende Merkmalsauswahlverfahren als auch einfach zu realisierende Klassifikationsverfahren bekannt. Es muß im Anwendungsfall entschieden werden, ob man mit einem leicht realisierbaren Merkmalsauswahlverfahren arbeiten muß und dann eben einen möglichst wirkungsvollen Klassifikator zu suchen hätte, oder ob es erforderlich ist, einen festen, einfachen Klassifikator zu verwenden, der dann eben mit geeigneten Merkmalen versorgt werden muß.

Für Konzept a) ist die Theorie weiter entwickelt als für Konzept b). Insbesondere kann man z. B. über die Hierarchie von Klassifikatoren aus Bild 2-2 eine Reihe von in gewissem Sinn geordneten Klassifikationsschemata einsetzen: Man teste zunächst mit dem einfachen MAK, bei ungenügender Leistung kann man über den MHK, MMA und NVK sich im Extremfall bis zum Bayesklassifikator durcharbeiten.

Konzept b) wird in der Literatur eigentlich nicht verfolgt. Es ist aber eine überlegenswerte Alternative zu Konzept a). Formal bedeutet Forderung b), daß Merkmale mit einer Gütefunktion zu bewerten sind, die sich aus der Struktur des verwendeten Klassifikators ableiten läßt, oder, mit anderen Worten, daß Merkmale mit einer Gütefunktion zu bewerten sind, aus deren Werte Rückschlüsse auf die Leistung des Systems bei Verwendung des vorgegebenen Klassifikators gezogen werden können. Es ist auch denkbar, daß die Vorgabe des Klassifikators als Hilfsmittel zur Auswahl einer Gütefunktion aus einer Tabelle, wie etwa Tabelle 2-1, dient.

Im Rahmen der vorliegenden Arbeit wird untersucht, wie mit Konzept b) gearbeitet werden kann. Die Resultate der Versuche zeigen, daß zum einen das Konzept überhaupt Bedeutung hat, zum anderen durch Anwendung des Konzeptes tatsächlich

geeignete Merkmale gefunden werden können. Die Art der Einführung des Konzeptes hat den Namen

"Klassifikatorbezogene Merkmalsauswahl"

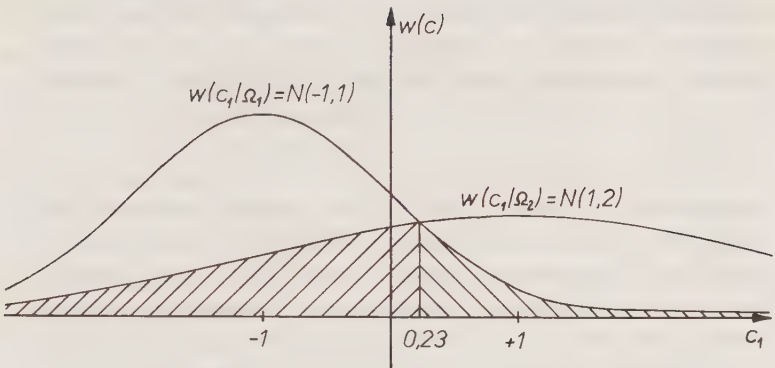
für diese Art der Merkmalsauswahl bzw. die Art der Bestimmung einer Merkmalsauswahlfunktion geeignet erscheinen lassen.

Das Konzept soll anhand eines einfachen Beispiels verdeutlicht werden. Es darf nicht übersehen werden, daß ein einfaches Beispiel in der Mustererkennung notgedrungen konstruiert wirkt. Es ist aber trotzdem oder gerade deswegen möglich, wesentliche Eigenschaften mit solchen Beispielen zu zeigen.

In einem hypothetischen Zwei-Klassen-Problem können an einem Muster die beiden Eigenschaften c_1 und c_2 gemessen werden. Bild 3-1 zeigt die klassenbedingten Dichten, die hier angenommen werden.

Merkmal c_1 ist für beide Klassen normalverteilt, wobei Mittelwerte und Streuungen unterschiedlich sind. Merkmal c_2 jedoch zeichnet sich dadurch aus, daß Messungen grundsätzlich um einen Wert μ schwanken. Muster der Klasse Ω_1 tendieren dazu, für c_2 Meßwerte zu liefern, die kleiner als μ sind, Muster der Klasse Ω_2 solche Meßwerte, die größer als μ sind. c_2 ist für keine der beiden Klassen normalverteilt. Im Beispiel wurden für die Dichten Chi-Quadrat-Verteilungen angenommen.

Es sei nun gefordert, daß nur eines der beiden Merkmale zur Klassifikation herangezogen wird. Der Einfachheit halber werden hier gleiche a priori Wahrscheinlichkeiten vorausgesetzt. Fordert man nun, daß der Minimum-Abstand-Klassifikator eingesetzt wird, so bleibt die Aufgabe zu lösen, c_1 und c_2 auf ihre Eignung als Eingangsgrößen für den MAK zu bewerten. Hierzu wird etwa mit Hilfe einer Teststichprobe

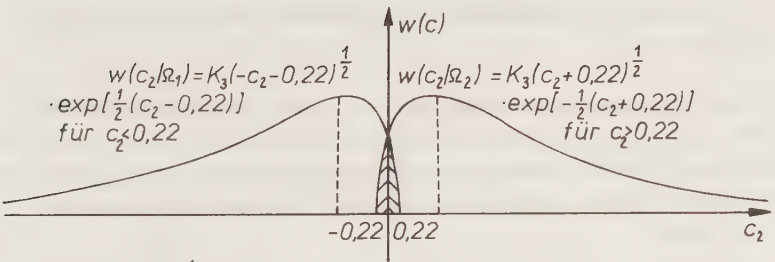


Fehlerwahrscheinlichkeit NVK: $p_f(\Omega_1) = 0,1144$

$p_f(\Omega_2) = 0,3519$

Fehlerwahrscheinlichkeit MAK: $p_f(\Omega_1) = 0,1587$

$p_f(\Omega_2) = 0,3085$



$K_3 = (2\pi)^{-\frac{1}{2}} = 0,3989$

Fehlerwahrscheinlichkeit (MAK und NVK): 0,05

Bild 3-1 Beispiel zur klassifikatorbezogenen Merkmalsauswahl

die zu erwartende Fehlerwahrscheinlichkeit ermittelt. Die schraffierten Flächen in Bild 3-1 geben ein Maß der Fehlerwahrscheinlichkeit bei Verwendung des jeweiligen Merkmals. Bei der angenommenen Lage der Dichten wird diese Prozedur also c_2 als zu verwendendes Merkmal auswählen.

Wird gefordert, daß der Normalverteilungsklassifikator (NVK) verwendet wird, so ist eine mögliche Auswahlprozedur, die vorhandenen Merkmale mit Hilfe statistischer Tests daraufhin zu prüfen, ob sie normalverteilt sind. Dies würde ganz im Sinne der Forderung aus Abschnitt 2.1.1 sein, daß vereinfachende Annahmen, die zur Verwendung spezieller Klassifikatoren führen, im konkreten Fall stets nachzuprüfen sind. Diese Prozedur würde c_2 immer als ungeeignet für den NVK ansehen und somit immer c_1 bevorzugen. Es ist aber nicht ausgeschlossen, daß c_2 auch bei Verwendung des NVK bessere Klassifikationsergebnisse bringen könnte als c_1 . Dies ist im gezeichneten Beispiel der Fall, die Fehlerraten sind in Bild 3-1 angegeben. Diese Effekte können am Beispiel natürlich gut demonstriert werden, sind in hochdimensionalen Merkmalräumen aber nicht mehr nachvollziehbar.

3.2 Klassifikatorbezogene Merkmalsauswahl beim Bayesklassifikator

Es soll nun gezeigt werden, wie mit dem Konzept "Klassifikatorbezogene Merkmalsauswahl" gearbeitet werden kann. Es wird hier zunächst ein formaler Ansatz dargestellt. In Kapitel 4 wird dann weiter darauf eingegangen, wie man eine Realisierung erhält.

Das Ziel jeder Merkmalsauswahl schlechthin wäre natürlich, solch einen Satz von Merkmalen zu finden, der das Bayesrisiko minimiert /F5/. Dieses Ziel ist aber nach dem bisher Gesagten vorsichtig zu interpretieren. Man wird häufig den Bayesklassifikator nicht zur Verfügung haben. Deshalb ist die Mini-

mierung des Bayesrisiko im Einzelfall möglicherweise nutzlos, eben weil man den Bayesklassifikator nicht verwendet. Weiterhin wird die Berechnung solcher optimalen Merkmale im Bayes-Sinn in praktischen Fällen nur dann möglich sein, wenn man vereinfachende Annahmen verwendet. Es bleibt dann wieder die Aufgabe, nachzuprüfen, ob diese Annahmen im konkreten Fall zulässig sind.

Der allgemeinste Ansatz zur Merkmalsauswahl durch Minimierung des Bayes-Risiko geht davon aus, daß k Klassen Ω_k vorgegeben sind. Zu jedem Muster wird ein n -dimensionaler Merkmalvektor $\underline{c}$ bestimmt. Die klassenbedingten Dichten $w(\underline{c}|\Omega_k)$, die a priori Wahrscheinlichkeiten $p(\Omega_k)$ und die Kosten $r(\Omega_k, \Omega_\lambda)$ seien bekannt. Unter diesen Voraussetzungen ist das Bayes-Risiko durch

$$V = \sum_{\lambda=1}^k \int_{\Omega_k^0} u_\lambda(\underline{c}) d\underline{c} \quad (3-1)$$

mit

$$u_\lambda(\underline{c}) = \sum_{\kappa=1}^k r(\Omega_\kappa, \Omega_\lambda) w(\underline{c}|\Omega_\kappa) p(\Omega_\kappa) \quad (3-2)$$

gegeben. Ω_k^0 ist nach Gleichung (2-5) das Gebiet des $\mathbb{R}^n$, das durch den Bayes-Klassifikator der Klasse Ω_k zugeordnet wird.

Sei nun $\underline{A}$ die Merkmalsextraktionsfunktion, die ein $\underline{c}$ aus dem $\mathbb{R}^n$ in ein $\underline{c}'$ aus dem $\mathbb{R}^{n'}$ überführt:

$$\begin{aligned} \underline{A}: \mathbb{R}^n &\rightarrow \mathbb{R}^{n'} \\ n' &< n \\ \underline{c}' &= \underline{A}(\underline{c}) \end{aligned} \quad (3-3)$$

Formal kann nun das Bayes-Risiko angeschrieben werden, falls man $\underline{c}'$ zur Klassifikation heranzieht. Wegen der Definition

von $\underline{c}'$ gemäß (3-3) ist dieses Risiko natürlich eine Funktion der Merkmalsextraktionsfunktion $\underline{A}$.

$$V(\underline{A}) = \sum_{\lambda=1}^k \int_{\Omega_{\lambda}^{O'}} u'_{\lambda}(\underline{c}') d\underline{c}' \quad (3-4)$$

mit

$$u'_{\lambda}(\underline{c}') = \sum_{\kappa=1}^k r(\Omega_{\kappa}, \Omega_{\lambda}) w(\underline{c}' | \Omega_{\kappa}) p(\Omega_{\kappa}) \quad (3-5)$$

Das Integral in (3-4) geht nun über das Gebiet $\Omega_{\lambda}^{O'}$ im $\mathbb{R}^{n'}$. Diese Teilmengen des $\mathbb{R}^{n'}$ sind natürlich Funktionen von $\underline{A}$. Die wesentliche Arbeit im Übergang von (3-3) nach (3-4) steckt jedoch in der Berechnung der Dichten $w(\underline{c}' | \Omega_{\kappa})$ aus $w(\underline{c} | \Omega_{\kappa})$. Liegen diese Dichten vor, folgen die Gebiete $\Omega_{\lambda}^{O'}$ aus der Definition des Bayes-Klassifikators.

Der Entwurf der Merkmalsauswahlstufe reduziert sich somit auf die Bestimmung der Funktion $\underline{A}^*$, die $V(\underline{A})$ aus (3-4) minimiert. Die so gestellte Aufgabe ist sehr allgemein. Man hat eine formal gestellte Optimierungsaufgabe, die das Bayes-Risiko im Zielraum als Zielfunktion verwendet. Bezüglich der Merkmalsextraktionsfunktion werden keine weiteren Voraussetzungen gemacht. $\underline{A}$ könnte also eine beliebige Menge von Funktionen durchlaufen. Da die Auswahl von n' aus n Merkmalen als Matrizenmultiplikation geschrieben werden kann, ist sie somit in die hier angestellten Betrachtungen mit eingeschlossen.

In gewisser Hinsicht hat man auf diese Weise wieder den Zustand erreicht, daß Merkmalsauswahl als Optimierungsaufgabe formuliert ist. Dies gilt ja auch für die Verfahren nach Abschnitt 2.2. Im Unterschied zu diesem Verfahren ist das "Gütemaß" in (3-4) aber nicht heuristisch konstruiert, sondern aus der Tatsache abgeleitet, daß der Bayes-Klassifikator verwendet werden soll und dieser ein Klassifikations-

risiko gemäß (3-1) und (3-4) aufweist. Gerade dies steht hinter dem Konzept der "Klassifikator-bezogenen Merkmalsauswahl".

Der allgemeine Ansatz nach Gleichung (3-3) und (3-4) führt sehr schnell auf erhebliche Schwierigkeiten in der analytischen Behandlung und auf unlösbare Probleme bei der praktischen Bestimmung der optimalen Auswahlfunktion $\underline{A}^*$. Selbst stark vereinfachende Annahmen wie klassenweise normalverteilte Merkmalvektoren, lineare Auswahlfunktionen $\underline{A}$ und Minimierung der Bayes-Fehlerrate führen auf schwer handhabbare numerische Probleme. Aspekte dieser Art werden in Kapitel 4 noch zu behandeln sein.

Es gibt Aussagen über die Existenz optimaler $\underline{A}^*/F5/$. Diese Aussagen kommen aus der Mathematik und besitzen primär theoretischen Wert, da sie vor allem auch Voraussetzungen über die Art der möglichen Funktionen $\underline{A}$ beinhalten. Generell wird angenommen, daß die Funktionen $\underline{A}$ aus einer parametrischen Familie stammen

$$\underline{A}(\underline{c}) = \Phi(\underline{c}, \underline{a}) \quad (3-6)$$

wobei $\underline{a}$ ein Parametervektor ist. Aus (3-4) wird also die Optimierungsaufgabe

$$V(\underline{A}) = V(\underline{a}) \stackrel{!}{=} \min \quad (3-7)$$

abgeleitet. In /F5/ werden vier Voraussetzungen angegeben, die sicherstellen, daß (3-7) lokale Minima haben kann. Diese Voraussetzungen, die in /F5/ rein mathematisch formuliert werden, werden in /J1/ aus der Sicht der Mustererkennung beleuchtet. Zwei dieser Voraussetzungen fordern die Stetigkeit der zweiten Ableitungen

$$\frac{\partial^2}{\partial c_i \partial c_j} \Phi(\underline{c}, \underline{a}) \quad , \quad i, j \in \{1, \dots, n\}$$

und

$$\frac{\partial^2}{\partial a_i \partial a_j} \underline{\phi}(\underline{c}, \underline{a}) .$$

Eine weitere Voraussetzung fordert, daß die Einschränkung von $\underline{\phi}(\underline{c}, \underline{a})$ auf gewisse $(n-n')$ -dimensionale Unterräume des $\mathbb{R}^n$ bijektiv ist. Diese Forderung soll sicherstellen, daß die Dichten $w(\underline{c}' | \Omega_k)$ existieren. Es genügt aber auch die Voraussetzung, daß jeder Punkt im Zielraum $\mathbb{R}^{n'}$ bezüglich der eingeschränkten Funktion $\underline{\phi}(\underline{c}, \underline{a})$ endlich viele Urbilder in diesem $(n-n')$ -dimensionalen Unterraum des $\mathbb{R}^n$ besitzt. Selbst in diesem Fall bleiben die Aussagen in /F5/ gültig /J1/. Bereits die bisher genannten Voraussetzungen reichen aus, daß die üblichen Bedingungen für die Existenz eines lokalen Minimums

$$\nabla_{\underline{a}} V(\underline{a}) \Big|_{\underline{a}^*} = \emptyset \quad (3-8)$$

$$\nabla_{\underline{a}} \nabla_{\underline{a}, t} V(\underline{a}) \Big|_{\underline{a}^*} \quad \text{positiv definit} \quad (3-9)$$

notwendig und hinreichend sind. In /F5/ wird als vierte Voraussetzung gefordert, daß die Funktion

$$g(\underline{c}) = u_{\lambda}(\underline{c}) - u_{\kappa}(\underline{c}) , \quad (3-10)$$

welche auf der Klassengrenze definitionsgemäß verschwindet, an der Klassengrenze zwischen Ω_{λ} und Ω_{κ} keine mehrfache Nullstelle besitzt. Dies kommt von der Seite der Mustererkennung der Forderung gleich, daß Entscheidungen auch in der Nähe der Klassengrenzen sicher getroffen werden können.

Die so gestellte Optimierungsaufgabe (3-7) kann bisher in dieser allgemeinen Form nicht auf praktische Probleme angewendet werden. Es ist nötig, die vorgestellten Ideen durch vereinfachende Annahmen in lösbarer Aufgaben zu transformieren. Die bereits mehrmals geäußerten Vorbehalte gegen vereinfachende Annahmen gelten auch hier in vollem Umfang.

Im Rahmen der vorliegenden Arbeit wurde das vorgestellte Konzept für klassenweise normalverteilte Merkmale und lineare Auswahlfunktionen zur Minimierung der Bayes-Fehler-rate implementiert. Die dabei auftretenden Schwierigkeiten und einige Resultate finden sich in Kapitel 4.

3.3 Klassifikatorbezogene Merkmalsauswahl beim modifizierten Minimum-Abstand-Klassifikator

Der modifizierte Minimum-Abstand-Klassifikator (MMA) wurde in Abschnitt 2.1.2.2 als Spezialfall des Bayes-Klassifikators eingeführt. Seine Prüfgrößen

$$u_k(\underline{c}) = (\underline{c} - \underline{\mu}_k) \underline{K}_k^{-1} (\underline{c} - \underline{\mu}_k) \quad (3-11)$$

können als Abstandskadrate des Merkmalvektors $\underline{c}$ zu den klassenspezifischen Mittelvektoren aufgefaßt werden, wobei jeder Klasse ein eigenes Abstandsmaß $d_k(.,.)$ zugeordnet wird. Man erhält das klassenspezifische Abstandsmaß, wenn man die Inverse der Kovarianzmatrix $\underline{K}_k$ als Kernmatrix der Bilinearform

$$(\underline{c}, \underline{c}') = \underline{c} \underline{A}_k \underline{c}' \quad (3-12)$$

verwendet. Jeder Ausdruck von der Form (3-12) liefert bekanntlich für positiv definite Matrizen $\underline{A}$ ein inneres Produkt auf dem zugrunde liegenden Vektorraum $/Z1/$. Kovarianzmatrizen sind positiv definite Matrizen, ebenso ihre Inversen. Weiterhin ist bekannt, daß für jedes innere Produkt $(.,.)$ die Definition

$$d(.,.) = \sqrt{(.,.)}$$

eine Metrik, also einen Abstand, liefert $/M2/$. Da die Klassifikation auf einem Größenvergleich der k Prüfgrößen basiert, ist eine monotone Abbildung der $u_k(\underline{c})$ für die Klassifikations-

leistung unerheblich. Deshalb kann in (3-11) das Abstandsquadrat eingesetzt werden.

Ziel ist nun, für diesen Klassifikator geeignete Merkmale zu finden. Der erste Schritt hierzu ist die Ableitung eines Gütemaßes aus dem Aufbau des Klassifikators. Daran anschließend ist dieses Gütemaß als Funktion der Merkmalsauswahlfunktion auszudrücken. Über die Auswertung und Optimierung der so gewonnenen Gütefunktion wird in Kapitel 4 berichtet.

Die grundlegende Überlegung zur Ableitung des Gütemaßes muß sicherlich darauf eingehen, wann der MMA Muster richtig bzw. falsch klassifiziert. Solche Betrachtungen sind für abstandsmessende Klassifikatoren deshalb relativ leicht durchzuführen, weil geometrische Anschauung gut eingesetzt werden kann. Prinzipiell wird ein Muster aus Klasse Ω_k durch den hier vorgegebenen Klassifikator dann in die richtige Klasse eingeordnet, wenn der Abstand des zugehörigen Merkmalvektors $\underline{c}$ zum Klassenzentrum $\underline{\mu}_k$ geringer ist als seine Abstände zu allen anderen Klassen. Man muß hier natürlich davon ausgehen, daß die Klassenzentren alle verschieden sind.

Es gilt nun, diese Aussage für ein Muster auf möglichst grose, leicht charakterisierbare Gruppen von Mustern auszudehnen. Zu diesem Zweck werden die Bereiche im Merkmalsraum betrachtet, die durch den Klassifikator einer Klasse zugeordnet werden. Diese Bereiche sind durch irgendwelche (Trenn-) Flächen zweiter Ordnung begrenzt. Solche Gebiete sind allerdings nicht einfach charakterisierbar. Man kann sich aber damit behelfen, eine Schar von einfach zu charakterisierenden Gebieten anzugeben und aus dieser Schar dasjenige auszuwählen, das mit Sicherheit noch in die vom Klassifikator induzierten Klassenbereiche fällt. In der Geometrie kennt man als solche Gebilde die "einbeschriebenen" Kreise, Rechtecke, Vielecke und andere regelmäßige Figuren. Für den hier betrachteten Fall kommt als Schar von einfach

zu charakterisierenden Gebieten die Menge aller Flächen mit

$$(\underline{c} - \underline{\mu}_k) t_{\underline{K}_k}^{-1} (\underline{c} - \underline{\mu}_k) = \text{const} \quad (3-13)$$

in Frage. Im Sinne der Geometrie handelt es sich um eine Schar von Hyperkugeln mit gemeinsamem Mittelpunkt. Die Darstellung im $\mathbb{R}^n$ erfolgt zweckmäßigerweise als Schar von konzentrischen Hyperellipsoiden.

Unter den so definierten Hyperellipsoiden sucht man nun dasjenige heraus, das gerade noch in das vom MMA induzierte Klassengebiet paßt. Alle Muster aus Klasse Ω_k , die innerhalb des so definierten Ellipsoids liegen, werden mit Sicherheit richtig klassifiziert. Ein Bezug zur Klassifikationsleistung des Gesamtsystems gelingt sicherlich, wenn der Anteil der Muster, die in diesem Ellipsoid zu liegen kommen, abgeschätzt werden kann. Das Problem bei der Festlegung der so eingeführten Ellipsoide liegt in der Bestimmung der Konstanten "const" in (3-13). Zunächst ist nur eine Optimierungsaufgabe gegeben, deren Lösung diese Konstante wird. Zur Formulierung dieser Optimierungsaufgabe betrachte man zunächst nur die zwei Klassen Ω_k und Ω_λ . Die Grenze zwischen diesen beiden Klassen ist definiert durch

$$(\underline{c} - \underline{\mu}_k) t_{\underline{K}_k}^{-1} (\underline{c} - \underline{\mu}_k) = (\underline{c} - \underline{\mu}_\lambda) t_{\underline{K}_\lambda}^{-1} (\underline{c} - \underline{\mu}_\lambda) \quad (3-14)$$

Die Konstante $R_{k\lambda}$, die den "Radius" des in das Klassengebiet der Klasse Ω_k einbeschriebenen Ellipsoides angibt, ist nichts anderes als der Minimalabstand aller Punkte auf der Klassengrenze nach (3-14) zum Klassenzentrum $\underline{\mu}_k$, gemessen in der passenden Metrik. In mathematisch präziser Form schreibt sich dies

$$R_{k\lambda}^2 = \min_{\underline{c}} (\underline{c} - \underline{\mu}_k) t_{\underline{K}_k}^{-1} (\underline{c} - \underline{\mu}_k) \quad (3-15)$$

unter der Nebenbedingung (3-14)

oder formuliert als Optimierungsaufgabe mit Zielfunktion und Restriktion

$$(\underline{c} - \underline{\mu}_k) \underline{t}_k^{-1} (\underline{c} - \underline{\mu}_k) \stackrel{!}{=} \text{Min} \quad (3-16)$$

$$(\underline{c} - \underline{\mu}_k) \underline{t}_k^{-1} (\underline{c} - \underline{\mu}_k) = (\underline{c} - \underline{\mu}_\lambda) \underline{t}_\lambda^{-1} (\underline{c} - \underline{\mu}_\lambda)$$

Einen Eindruck von solchen Ellipsoiden gibt Bild 3-2. Die Ellipsoide im dargestellten $\mathbb{R}^2$ sind Ellipsen, die Klassengrenzen Kegelschnitte.

(3-16) ist eine Aufgabe der nichtlinearen Programmierung. Verfahren zur Lösung solcher Aufgaben sind bekannt /L1, H3/. Ein schnell konvergierendes Verfahren für die Implementierung wird in Abschnitt 4.4.3.2 kurz dargestellt.

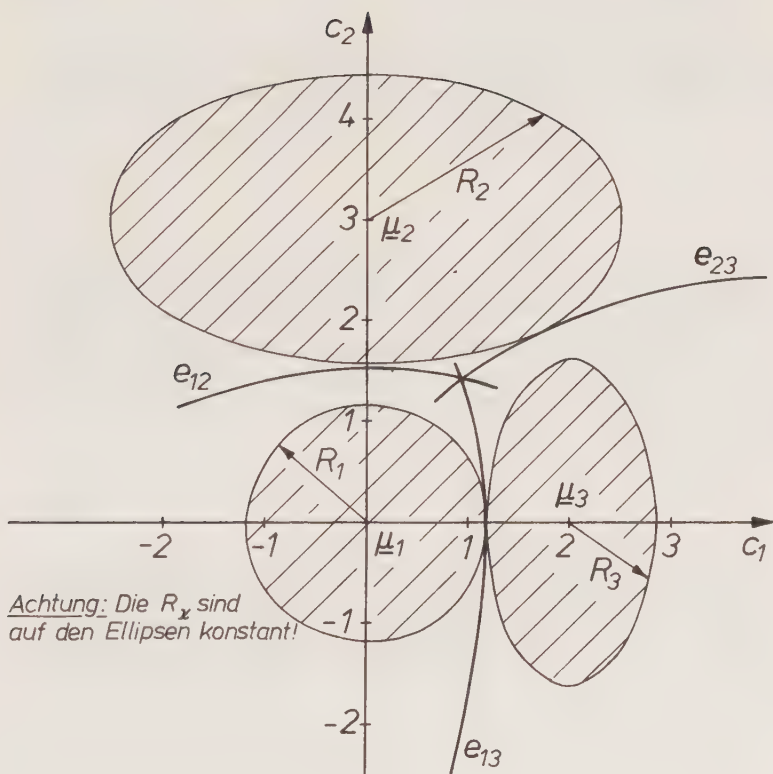
Die Erweiterung des Konzepts auf k Klassen ist leicht möglich. Man definiere für jede Klasse den Radius R_k des in das Klassengebiet eingeschriebenen Ellipsoides durch

$$R_k = \min_{\lambda} R_{k\lambda} \quad (3-17)$$

Intuitiv sind als Eingangsgrößen für den MMA solche Merkmale besonders gut geeignet, für die die Größen R_k besonders groß sind. Die eigentliche Forderung müßte natürlich sein, daß die Ellipsoide mit R_k als Radius möglichst viele Muster der Klasse Ω_k umschließen, d. h. möglichst viel Wahrscheinlichkeitsmasse der klassenbedingten Dichten enthalten. Es ist leicht möglich, aus den Radien R_k eine ganze Palette von Gütefunktionen für Merkmale, auch im Falle von k Klassen, abzuleiten. Beispiele sind etwa

$$G = \min_k R_k \quad (3-18)$$

ein Maß, daß das kleinste Ellipsoid aufsucht, oder



$$\Omega_1: \mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}; \quad K_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}; \quad R_1^2 = 1,372$$

$$\Omega_2: \mu_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}; \quad K_2 = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}; \quad R_2^2 = 2,074$$

$$\Omega_3: \mu_3 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}; \quad K_3 = \begin{pmatrix} 0,5 & 0 \\ 0 & 2 \end{pmatrix}; \quad R_3^2 = 1,372$$

Bild 3-2 In Klassengebiete eingeschriebene Ellipsen

$$G = \frac{1}{k} \sum_{\kappa=1}^k R_{\kappa} \quad (3-19)$$

als einfaches arithmetisches Mittel der Radien. Es ist auch denkbar, ein Gütemaß zu definieren, das eine Gewichtung mit a priori Wahrscheinlichkeiten enthält

$$G = \sum_{\kappa=1}^k p(\Omega_{\kappa}) \cdot R_{\kappa} \quad (3-20)$$

Ein für den hier angestrebten Zweck besonders günstiges Gütemaß läßt sich aus den Zahlen R_{κ} ableiten, wenn man bedenkt, daß es sich bei dem Ausdruck (3-13) um eine quadratische Form handelt, da $\underline{K}_{\kappa}^{-1}$ positiv definit ist, sogar um eine elliptische Form. Betrachtet man nun wieder $\underline{c}$ als Zufallsvariable, so ist auch

$$Q_{\kappa} = (\underline{c} - \underline{\mu}_{\kappa}) \underline{K}_{\kappa}^{-1} (\underline{c} - \underline{\mu}_{\kappa}) \quad (3-21)$$

eine Zufallsvariable. Für solche quadratischen Formen gibt es eine Erweiterung der Tschebyscheff'schen Ungleichung aus der Statistik. Es gilt nämlich unabhängig von den zugrundeliegenden Verteilungen /W3/

$$P(Q_{\kappa} \geq \alpha^2) < \frac{n}{\alpha^2} \quad (3-22)$$

Dabei ist n die Dimension des Merkmalraumes. Aus dieser Ungleichung läßt sich sofort eine Fehlerschranke bei Verwendung des MMA ableiten. Setzt man nämlich α mit R_{κ} gleich, so bedeutet dies nichts anderes, als daß man sich für die Wahrscheinlichkeit interessiert, daß ein Muster aus Klasse Ω_{κ} außerhalb des Ellipsoids liegt, das die mit Sicherheit richtig klassifizierten Muster enthält. Alle Muster außerhalb des Ellipsoids mit R_{κ} als Radius werden als Kandidaten für eine Fehlklassifikation angesehen. Da es aber auch Muster geben kann, deren Merkmalsvektor außerhalb des Ellip-

soides mit R_k als Radius und μ_k als Mittelpunkt liegen kann und die trotzdem mit Hilfe des MMA korrekt klassifiziert werden können, muß die echte Fehlerwahrscheinlichkeit kleiner sein als $P(Q_k > R_k^2)$. Damit erhält man aus Gleichung (3-22) als Abschätzung für die klassenbedingte Fehlerwahrscheinlichkeit /01/

$$P_f(\Omega_k) < \frac{n}{R_k^2} \quad (3-23)$$

Diese Schranke ist wegen ihrer Ableitung aus der Tschebyscheff-Ungleichung natürlich sehr grob. Sie liefert aber in jedem Fall eine Reihe weiterer Gütekriterien, die mit der Leistung des MMA in Zusammenhang stehen. Naheliegende Zahlen sind etwa

$$G = \max_k \frac{n}{R_k^2} \quad (3-24)$$

also der maximale Klassifikationsfehler bei Mustern aus einer Klasse oder

$$G = \sum_k P(\Omega_k) \cdot \frac{n}{R_k^2} \quad (3-25)$$

eine Schätzung für den Klassifikationsfehler überhaupt. In beiden Fällen wird als Klassifikator der MMA gefordert. Beide Gütemaße sollten möglichst kleine Werte annehmen. Es werden bei diesen Gütemaßen insbesondere solche Merkmale als gänzlich ungeeignet für den MMA bewertet, die Klassenmittelpunkte zusammenfallen lassen. In diesem Fall ergeben sich die R_k als Null! Eine bessere Schranke für die Fehlerwahrscheinlichkeit ergibt sich, wenn man klassenweise normalverteilte Merkmalvektoren annimmt /01/. In diesem Fall erhält man

$$P_f(\Omega_k) < \frac{\Gamma\left(\frac{n}{2}, \frac{R_k^2}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \quad (3-26)$$

eine monotone Funktion von R_k . Dabei ist

$$\Gamma(\alpha, x) = \int_x^{\infty} t^{\alpha-1} e^{-t} dt$$

die unvollständige Gammafunktion und

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt$$

die gewöhnliche Gammafunktion. Die Anwendung dieser Schranke setzt allerdings wieder voraus, daß im konkreten Problem tatsächlich klassenweise normalverteilte Merkmalvektoren vorliegen.

Der Schlüssel zur klassifikatorbezogenen Merkmalsauswahl liegt in jedem Fall in der Berechnung der Zahlen R_k . Diese Berechnung ist im Zielraum der Merkmalsextraktionsfunktion durchzuführen. Anschließend ist eines der angegebenen Gütekriterien auszuwerten. Für die Berechnungen gilt das in 3.2 für die klassifikatorbezogene Merkmalsauswahl beim Bayes-klassifikator Gesagte: die analytische Behandlung schon der Aufgabe (3-16) wird nahezu unmöglich, bei stark vereinfachenden Annahmen gelingt es, numerische Lösungswege zu finden. Für lineare Merkmalsextraktionsfunktionen

$$\underline{c}' = \underline{A}_t \underline{c} \quad (3-27)$$

wird aus (3-11)

$$u_k(\underline{c}') = (\underline{c} - \underline{\mu}_k)_t \underline{A}(\underline{A}_t \underline{K}_k \underline{A})^{-1} \underline{A}_t (\underline{c} - \underline{\mu}_k) \quad (3-28)$$

ein Ausdruck, der bezüglich $\underline{A}$ zu maximieren ist, wenn man etwa das Gütemaß (3-18) für die Auswahl zugrunde legt.

Im Rahmen dieser Arbeit wird in Kapitel 4 eine Realisierung des Auswahlverfahrens beschrieben, das für lineare Merkmalsextraktionsfunktionen das Gütemaß nach (3-25) minimiert.

3.4 Verallgemeinerung des Konzepts für andere Klassifikatoren

Ein Konzept wie das der klassifikatorbezogenen Merkmalsauswahl ist vor allem dann interessant, wenn es gelingt, seine Anwendung in möglichst vielen Fällen zu garantieren. Die Forderung des Konzepts ist klar: zur Bewertung von Merkmalen bei der Merkmalsauswahl sollen nur solche Gütemaße verwendet werden, die einen Bezug zu dem Klassifikator herstellen lassen, der im Endeffekt für die Klassifikationsstufe verwendet wird. Im Falle der Verwendung des Bayes-Klassifikators wurde gezeigt, daß tatsächlich das Bayes-Risiko als Funktion der Merkmalsauswahlfunktion geschrieben werden kann. Für den MMA wurde gezeigt, wie mit Hilfe von Hyperellipsoiden die Klassifikationsleistung gerade des MMA charakterisiert werden kann. Die für den MMA verwendete Methode der einbeschriebenen Hyperellipsoide kann im Prinzip für alle abstandsmessenden Klassifikatoren verwendet werden. Der Minimum-Abstand-Klassifikator erhält insofern eine einfache Struktur, als alle Kernmatrizen für die Bilinearformen (3-12) die Einheitsmatrix werden. Für diesen Fall ist jedoch eine Abschätzung der Fehlerrate über die Tschebyscheff-Ungleichung nicht mehr möglich, da diese gerade die inversen Kovarianzmatrizen als Kernmatrizen verlangt.

Die Anwendung des Konzepts auf zum Beispiel geometrische Klassifikatoren oder die Nächste-Nachbar-Regel erfordert eine weitergehende Untersuchung dieser Klassifikatortypen. Die Nächste-Nachbar-Regel hat zwar einige bemerkenswerte theoretische Eigenschaften, man darf aber nicht übersehen, daß bereits ihre Anwendung, das heißt also die Verwendung dieser Regel als Klassifikator, mit relativ hohem Aufwand verbunden sein kann, wenn nur die Lernstichprobe umfangreich genug wird. Die Eigenschaften dieser Regel gelten aber nur für Lernstichproben mit potentiell unendlich vielen Elementen.

Die Anwendung des Konzepts auf analytischem Weg ist im all-

gemeinen Fall zumindest noch schwierig. Aus der Realisierung der Verfahren in Kapitel 4 wird sich aber ein Vorschlag ergeben, der unmittelbar auf beliebige Klassifikatoren übertragbar ist.

4. Experimentelle Untersuchungen

Mustererkennung als Wissenschaft versucht, intelligente Fähigkeiten von Lebewesen, insbesondere auch von Menschen, zu simulieren. Zu diesem Zweck werden Verfahren entworfen, die spezielle Teilaufgaben, wie zum Beispiel Schriftzeichen-erkennung, zu lösen gestatten. Es ist in vielen Fällen möglich, auf theoretischem Weg Aussagen über die Leistungsfähigkeit solcher Verfahren zu gewinnen. Solche Aussagen sind beispielsweise Formeln für Fehlerschranken oder die Feststellung, daß dieses oder jenes Verfahren "besser" ist als ein anderes. In den bislang erschienenen Lehrbüchern zur Mustererkennung, wie /F1, D1, N2/, werden naturgemäß solche Verfahren dargestellt, die auf solidem theoretischen Hintergrund beweisbare Aussagen liefern. Es hat sich im Laufe der letzten Jahre auch tatsächlich ein umfangreiches Wissen über den theoretischen Bereich der Mustererkennung angesammelt.

Diese Ausweitung und Aufbereitung des theoretischen Wissens über die Phänomene in der Mustererkennung darf jedoch nicht dazu führen, daß Mustererkennung als theoretisches Fach aufgefaßt wird. Von der Konzeption her handelt es sich bei der Mustererkennung um ein ingenieurmäßig zu betreibendes Gebiet. Es geht um die Konstruktion von neuen Verfahren und Geräten. Die Analyse vorgegebener Systeme, wie sie im Bereich der Naturwissenschaft eine Rolle spielt, muß zwar Bestandteil der Wissenschaft Mustererkennung sein, steht jedoch erst an zweiter Stelle.

Wenn die Mustererkennung neue Verfahren und Geräte konstruiert, so muß sie sich zwangsläufig auch der Aufgabe unterziehen, die Leistung dieser Verfahren und Geräte zu demonstrieren. Dies kann bis zu einem gewissen Grade natürlich auf theoretischem Wege geschehen. Letztlich unerlässlich bleibt aber das Experiment, das in der realen Situation die

Leistungsfähigkeit eines Verfahrens aus der Mustererkennung demonstriert. Solche Experimente haben sowohl die absolute Leistung eines Systems aufzuzeigen als auch die Leistung eines Verfahrens oder Systems in Relation zur Leistung anderer Verfahren oder Systeme zu setzen.

Ein Kernpunkt der vorliegenden Arbeit ist das Experiment mit sogenannten "realen Daten". Hinter dem Begriff "reale Daten" verbirgt sich die Idee, zum Experimentieren solche Muster zu verwenden, die in einem praktisch relevanten Teil der Umwelt tatsächlich auftreten. Dazu gehört auch, daß ein genügend großes Interesse an der automatischen Verarbeitung dieser Muster besteht. Solche Forderungen kommen zum Beispiel dann zustande, wenn in "kurzer" Zeit "viele" Muster zu verarbeiten sind. Beleglesemaschinen zum Beispiel sollen einige Tausend Schriftzeichen pro Sekunde verarbeiten können.

Zum Wesen einer Theorie gehört ihre Allgemeingültigkeit. Soll ein theoretisches Resultat in der Praxis, das heißt an realen Daten, nachgeprüft werden, so genügt es nicht, ein einziges Experiment durchzuführen und dessen Ergebnis als Kriterium für die Brauchbarkeit einer Theorie heranzuziehen. Vielmehr sind gerade im Bereich Mustererkennung viele Experimente notwendig, um überhaupt einen Einblick in die Gegebenheiten der realen Daten gewinnen zu können. Daran anschließend kann erst über Wert und Unwert eines Verfahrens oder einer Theorie geurteilt werden.

Die Experimente, die im Rahmen der vorliegenden Arbeit durchgeführt werden, sind unter den genannten Gesichtspunkten zu sehen. Es sind Experimente, die mit realen Daten durchgeführt wurden. Die Einbeziehung realer Daten führt sofort auf eine ganze Palette von Schwierigkeiten, die beim Bedarf an Rechnerkernzeit anfangen und bis hin zu Phänomenen reichen, für die wiederum im theoretischen Bereich Erklärungen gefunden werden müssen. Es soll hier ein Einblick in die Fülle der

Probleme und Phänomene gegeben werden, die die Durchführung von Experimenten mit realen Daten schwer werden lassen.

4.1 Probleme bei der Realisierung von Mustererkennungssystemen

Eine große Klasse von Problemen, die bei der Realisierung von Mustererkennungssystemen für reale Daten auftreten, kann darauf zurückgeführt werden, daß die klassenbedingten Dichten $w(\underline{c}|\Omega_k)$ des Merkmalvektors nicht bekannt sind. Als einziger Ersatz steht eine klassifizierte Stichprobe ω zur Verfügung. Diese enthält N Muster des Problemkreises. Es wird angenommen, daß von jeder Klasse Ω_k eine Anzahl N_k von Mustern in ω enthalten sind.

Diese endliche Stichprobe muß sowohl für die Extraktion der Klassencharakteristika ausreichen als auch für Experimente zur Leistungsmessung einer für das konkrete Problem entworfenen Klassifikationsprozedur herangezogen werden. Für statistische Klassifikatoren heißt das zunächst, daß die Dichten $w(\underline{c}|\Omega_k)$ zu schätzen sind, zumindest jedoch Parameter dieser Verteilungen.

Damit ist man bereits mit dem Problem der Parameterschätzung konfrontiert, ehe man überhaupt an typischere Mustererkennungsaufgaben wie Klassifikatoranpassung oder Merkmalsauswahl denken kann. Das Gebiet Parameterschätzung ist zwar theoretisch umfassend erforscht worden, die Forderungen an die Qualität und den Umfang von Stichproben sind jedoch enorm. Als ein für die Mustererkennung typisches Problem darf die Schätzung einer Kovarianzmatrix angesehen werden. Kovarianzmatrizen werden z. B. zur Dimensionierung des Normalverteilungsklassifikators und einiger seiner Spezialisierungen oder zur Bestimmung des Karhunen-Loève-Koordinatensystems benötigt. Von der bekannten Schätzformel

$$K = \frac{1}{N-1} \sum_{\rho=1}^N (\rho_{\underline{c}} - \underline{m}) (\rho_{\underline{c}} - \underline{m})_t \quad (4-1)$$

mit

$$\underline{m} = \frac{1}{N} \sum_{\rho=1}^N \rho_{\underline{c}} \quad (4-2)$$

weiß man zwar, daß sie eine erwartungstreue und konsistente Schätzung liefert; solche Aussagen gelten aber nur, wenn der Stichprobenumfang N beliebig groß gemacht werden kann /W3/. Interessanter sind Aussagen über die Genauigkeit einer Schätzung wie der aus (4-1), wenn man den endlichen Stichprobenumfang N verwendet. Aussagen solcher Natur liefern die Konfidenzintervalle oder Zuverlässigkeitsbereiche. Je kleiner ein Konfidenzintervall ist, desto zuverlässiger ist die zugehörige Schätzung, und Konfidenzintervalle werden umso kleiner, je größer der Stichprobenumfang wird. So werden für eine einigermaßen zuverlässige Schätzung einer Kovarianzmatrix einer Ordnung unter Hundert bereits einige Tausend Stichprobenelemente verlangt /N2/. Es kann sehr mühsam sein, solche Stichproben zu sammeln. So müßte zum Beispiel der Benutzer eines Kommandoeingabesystems jedes der erlaubten Kommandos einige zig-mal dem System vorsprechen und dann noch über einen zweiten Kanal, beispielsweise über eine Tastatur, die Bedeutung des Gesprochenen (= Klassenzugehörigkeit) dem System zweifelsfrei mitteilen.

Erstes und wichtigstes Problem jeder Mustererkennung ist - und wird in nächster Zukunft bleiben - die Stichprobe. Ohne eine repräsentative Stichprobe kann weder ein Klassifikationssystem dimensioniert werden noch können Aussagen über die Leistung eines solchen Systems gewonnen werden. Steht eine Stichprobe zur Verfügung, so ist über ihre Verwendung zu entscheiden. Es ist bekannt, daß Aussagen über die Leistung eines Systems optimistisch ausfallen, wenn zur Dimensionierung und zur Leistungsmessung des Systems die

gleiche Stichprobe herangezogen wird /T1/. Es stellt sich demnach die Frage nach der Aufteilung bzw. der besten Verwendung des zur Verfügung stehenden Datenmaterials. Die Frage wurde im Zusammenhang mit der Fehlerschätzung bei Klassifikationssystemen ausführlich untersucht, ein Überblick über die dazu erschienene Literatur findet sich in /T1/. Ein generell akzeptierter Vorschlag wird im Rahmen dieser Arbeit verwendet: die Gesamtstichprobe wird in eine Lernstichprobe und eine Teststichprobe aufgeteilt. Die Lernstichprobe wird zur Dimensionierung des Klassifikationssystems herangezogen, die Teststichprobe zur Leistungsmessung. Weitgehend ungeklärt bleibt das optimale Verhältnis von Umfang Lernstichprobe zu Umfang Teststichprobe und dessen Auswirkung auf die Genauigkeit der Resultate /H4/.

Ein weiteres Problem, das mit dem Problem der Schätzung von Verteilungsdichten zusammenhängt, ergibt sich, wenn man die Klassifikationsleistung eines Systems in Abhängigkeit von der Zahl der verwendeten Merkmale betrachtet. Man würde zunächst erwarten, daß die Leistung eines Systems mit der Zahl der verwendeten Merkmale zunimmt. Man kann entsprechende theoretische Aussagen auch ableiten, jedoch zeigt sich im Experiment, daß die Leistung eines Klassifikationssystems wieder abnimmt, wenn man die Anzahl der verwendeten Merkmale steigert, den Stichprobenumfang aber fest läßt /K1/. Es wurden einige Veröffentlichungen bekannt, die diese Phänomene zu erklären versuchen /H2, F7/. Die Autoren führen die Phänomene immer auf Schätzprobleme zurück. Es scheint jedoch so zu sein, daß eine Mischung mehrerer Effekte vorliegt. Als Haupteffekte dürfte tatsächlich das in /H2/ erwähnte Schätzproblem auftreten, es muß aber wohl auch berücksichtigt werden, daß bei großen Merkmalszahlen häufig Annahmen wie klas senweise Normalverteilungen einfach nicht mehr stimmen /O2/.

Im Rahmen der vorliegenden Arbeit wurde versucht, die angesprochenen Effekte soweit als irgendetmöglich klein zu halten.

Dies wurde durch die Verwendung möglichst weniger Merkmale erreicht. Zweck und Ziel der Arbeit soll die Darstellung und Erprobung von Merkmalauswahlverfahren sein. Die dazu nötigen Vergleiche können natürlich ohne weiteres auch im $\mathbb{R}^2$ durchgeführt werden. Man muß jedoch in jedem Fall dann nachprüfen, ob die gemessenen Relationen zwischen den Merkmalauswahlverfahren erhalten bleiben, wenn man zu höheren Dimensionen des Merkmalraumes übergeht. Es wurde versucht, auch in dieser Hinsicht Erkenntnisse zu gewinnen, indem Versuchsreihen mit der Dimension des Zielraumes als Parameter in die Experimente eingeschlossen wurden.

Die Beschränkung auf kleine Merkmalszahlen soll in keinem Fall heißen, daß auf diesem Weg wieder der Bezug zu den realen Daten eingeschränkt wird. In einigen Fällen wurden Merkmalszahlen so gewählt, daß "konkurrenzfähige" Klassifikationsleistungen erreicht wurden. Es soll hier darauf aufmerksam gemacht werden, daß die Einschränkung der Versuche auf kleine Merkmalszahlen auch rechentechnische Gründe hat. Es zeigt sich nämlich, daß der Entwurf einer Merkmalsextraktionsfunktion eine rechenintensive Aufgabe für eine Rechenanlage ist. Eine einzige Anpassung kann sehr schnell bis zu Wochen und Monaten an Rechenzeit erfordern. Bei den Versuchen im Rahmen der vorliegenden Arbeit war die hauptsächliche Größe, die die erforderliche Rechenzeit bestimmte, die Zieldimension der Auswahlfunktion. Zwischen Rechenzeit und Zieldimension besteht im wesentlichen eine quadratische Abhängigkeit. Die Beschränkung auf kleine Zieldimensionen machte überhaupt erst eine vergleichende Untersuchung verschiedener Verfahren möglich. Es wurde aber nicht versäumt, Rechenzeitschätzungen für die Bestimmung von Auswahlfunktionen mit hoher Zieldimension zu ermitteln, so daß im Einzelfall der erforderliche Aufwand für eine spezielle Berechnung angegeben werden kann. Die entsprechenden Tabellen finden sich in Abschnitt 4.5.

Aus praktischen Gründen ist es erforderlich, die Bestimmung der Dichten $w(\underline{c} | \Omega_K)$ auf eine Parameterschätzung zurückzuführen

ren. Dies bedeutet, daß ein Modell für die Dichten vorgegeben wird, wobei die Übereinstimmung des Modells mit der Wirklichkeit nachzuprüfen bleibt. Im Rahmen der vorliegenden Arbeit wird häufig angenommen, daß die Dichten $w(\underline{c}|\Omega_K)$ Dichten von Normalverteilungen sind. Dies reduziert die Aufgabe der Parameterschätzung auf die Bestimmung von Mittelvektoren nach Gleichung (4-2) und Kovarianzmatrizen nach Gleichung (4-1). Inwieweit die Annahme der Normalverteilung bei den einzelnen Datensätzen gerechtfertigt ist, wird festzustellen sein. In jedem Fall kann zunächst mit der Hypothese Normalverteilung gearbeitet werden. Man darf hier eben nicht erwarten, daß Klassifikationen wie NVK, MMA oder QMK (siehe Kapitel 2) optimale Klassifikatoren sind. Die Modellannahme Normalverteilung dient hier nur als Hilfsmittel, um zum Beispiel Hinweise für die zu berechnenden Parameter zu bekommen.

Zusammenfassend bleibt festzustellen, daß die hier durchzuführenden Experimente im Licht der geschilderten Probleme zu sehen sind. Es wurde versucht, die bekannten Fehlerquellen, Stichprobenumfang, Aufteilung der Stichprobe in Lern- und Teststichprobe und zuviele Merkmale für eine vorgegebene Stichprobe auszuschalten.

4.2 Datensätze

Die Experimente wurden mit zwei Stichproben durchgeführt, die beide ohne Einschränkung als wirklichkeitsnah bezeichnet werden können. Eine der beiden Stichproben besteht aus ca. 22400 geschriebenen Ziffern, jeweils ca. 2240 aus den 10 Klassen "0" bis "9". Die andere Stichprobe enthält ca. 15000 gesprochene Ziffern, jeweils ca. 1500 Stichprobenelemente aus jeder der 10 Klassen "0" bis "9".

Die Schriftzeichen liegen als binäre Rasterbilder in einem 32 x 40 Raster vor. Es handelt sich um Handblockschrift verschiedener Schreiber. Für die Zwecke der vorliegenden Arbeit

wurden die Zeichen einer Vorverarbeitung unterzogen. Dazu wurde das umschreibende Rechteck eines Zeichens in ein 12×16 Raster eingebettet. Nach dieser Prozedur lagen also alle Zeichen als Rasterbilder mit 12×16 binären Elementen vor.

Die Gesamtstichprobe wurde - mehr oder minder willkürlich - in eine Lernstichprobe mit ca. 10.000 Mustern und eine Teststichprobe mit ca. 12.000 Mustern unterteilt, so daß also zwei echt disjunkte Teilstichproben für die Durchführung der Experimente zur Verfügung standen.

Die Sprachdaten stammen von 10 verschiedenen Sprechern. Jeder Sprecher sprach alle Zahlworte an mehreren verschiedenen Tagen. Das Resultat der Aufnahme liegt in vorverarbeiteter Form als 14×20 Matrix aus reellen Zahlen vor. Den 14 Zeilen der Matrix entsprechen die Energieanteile jeweils eines zugeordneten Terz-Bandfilters, den 20 Spalten entspricht eine regelmäßige Unterteilung der Zeitdauer des gesprochenen Wortes in 20 gleichlange Zeitsegmente. Die Aufnahme-prozedur und Vorverarbeitung der Sprachdaten ist in /B1/ beschrieben.

Auch die Sprachstichprobe wurde in eine Lern- und eine Teststichprobe unterteilt. Die Lernstichprobe umfaßt wieder ca. 10.000 Muster, jeweils 1.000 Muster aus jeder Klasse. Die Teststichprobe enthält ca. 5.000 Muster, etwa 500 aus jeder Klasse.

Die so gegebenen Stichproben wurden jeweils mit der Karhunen-Loève-Transformation in einen 70-dimensionalen Merkmalsraum abgebildet. In beiden Fällen wurde das KL-Koordinatensystem aus der Kovarianzmatrix der Lernstichprobe ermittelt. Mit dieser Operation wurden drei Ziele angestrebt:

- i) Es wurde eine Dimensionsreduktion von 192 Merkmalen bei den Schriftzeichen bzw. 280 Merkmalen bei den Sprachdaten auf 70 Merkmale erreicht.

- ii) Die so gewonnenen Merkmale sind in beiden Fällen annähernd normalverteilt.
- iii) Durch die Reduktion gewinnt man in beiden Fällen eine einheitliche Repräsentation der Daten mit Hilfe eines 70-dimensionalen Merkmalvektors.

Die Punkte i) bis iii) machen die Vorabbildung der Muster sicher wünschenswert, es muß jedoch die Zulässigkeit der Operation begründet werden. Der Hauptgrund für die Zulässigkeit der Operation ist dadurch gegeben, daß die KL-Transformation als Reihenentwicklung nach einer Orthonormalbasis aufgefaßt werden kann, bei welcher der geometrische Rekonstruktionsfehler minimiert wird. Dieser Fehler ist sogar bekannt und kann leicht ermittelt werden /N2/. Er beträgt bei der Sprache 4,22 %, bei der Schrift 14 %. Dieser Rekonstruktionsfehler ist für die hier angestellten Untersuchungen insofern tragbar, als davon ausgegangen werden kann, daß durch die fehlerhafte Repräsentation der Muster keine Beeinträchtigung in den Klassifikationsleistungen zu erwarten ist.

Ein zweiter Grund für die Zulässigkeit der KL-Transformation als vorgezogene Merkmalsreduktion ist darin zu sehen, daß es sich um eine lineare Abbildung vom hochdimensionalen auf den niederdimensionalen Raum handelt. Die experimentellen Untersuchungen im Rahmen der vorliegenden Arbeit beschränken sich auf die Erprobung linearer Auswahlverfahren. Diese Auswahlverfahren werden ausgehend vom niederdimensionalen Raum ermittelt. Es ist leicht möglich, beide, hintereinander auszuführende Operationen zu einer einzigen linearen Abbildung zusammenzufassen, so daß die Aufteilung der Bestimmung der besten linearen Merkmale in zwei Teilschritte keinen Effizienzverlust für das Gesamtsystem bedeutet. Die Optimalitätseigenschaften der Gesamtabbildung ergeben sich dann aus der Tatsache der guten geometrischen Rekonstruierbarkeit im Zielraum der KL-Transformation und den Optimalitätseigenschaften der daran anschließenden linearen Abbildung.

Als letztes bedarf Punkt ii) einer Begründung. In /N2/ wird

festgestellt, daß die Entwicklungskoeffizienten der KL-Entwicklung als Summe von Zufallsvariablen gemäß dem zentralen Grenzwertsatz der Statistik annähernd normal verteilt sein sollten. Dieser Satz verlangt allerdings Summen, deren Gliederzahl beliebig groß wird. Auch wird die statistische Unabhängigkeit der Summanden verlangt. Beides ist hier nicht gegeben, so daß obige Behauptung mit Vorsicht zu betrachten ist. Experimentelle Untersuchungen haben aber die These in Punkt ii) unterstützt. So werden in /O2/ und /N3/ anhand einer Stichprobe von Schriftzeichen die Verteilungen der ersten KL-Koeffizienten mit dem Kolmogoroff-Smirnow-Test auf die These Normalverteilung untersucht und auch angenommen.

4.3 Das Experimentiersystem

Die experimentellen Untersuchungen wurden an der Rechenanlage TR440 des regionalen Rechenzentrums Erlangen durchgeführt. Es wurde ein Programmsystem aufgebaut, das verschiedene Untersuchungen in sehr flexibler Weise durchzuführen gestattet. Dazu gehört, daß elementare Operationen wie Schätzen einer Kovarianzmatrix zu einer Stichprobe, Berechnung von Eigenwerten und Eigenvektoren, lineare Transformation einer Stichprobe und ähnliche als eigenständige Programme mit definierter Schnittstelle auf der Ebene der Dateiverwaltung realisiert wurden. Durch dieses Konzept wurde erreicht, daß der Ablauf eines bestimmten Experiments durch eine Art Programmiersprache beschrieben werden konnte, die speziell auf die Zwecke der hier durchzuführenden Versuche zugeschnitten ist.

Das Versuchsprogramm enthielt Experimente mit drei Klassifikatoren und fünf verschiedenen Merkmalauswahlverfahren. Im einzelnen standen zur Verfügung:

A) Klassifikatoren

- i) der Normalverteilungsklassifikator (NVK)
(siehe Abschnitt 2.1.2.1)
- ii) der modifizierte Minimum-Abstand-Klassifikator (MMA)
(siehe Abschnitt 2.1.2.2)
- iii) der im quadratischen Mittel angepaßte Klassifikator
(QMK)
(siehe Abschnitt 2.1.3)

B) Merkmalsauswahlverfahren

- i) lineare Merkmalsauswahl nach der Karhunen-Loève-Transformation (KL-optimale Merkmale, siehe Abschnitt 2.2.2.3)
- ii) lineare Merkmalsauswahl durch Maximierung der mittleren Divergenz über alle Klassenpaare für klassenweise normalverteilte Merkmale (DIV-optimale Merkmale, siehe Abschnitt 2.2.1, insbesondere Tabelle 2-1)
- iii) lineare Merkmalsauswahl durch Minimierung der Fehlerate beim MMA (MMA-optimale Merkmale, siehe Abschnitt 3.3)
- iv) lineare Merkmalsauswahl durch Minimierung der Fehlerate beim NVK (NVK-optimale Merkmale, siehe Abschnitt 3.2)

Eine weitere Auswahlmethode, hier iterierte Hauptachsentransformation oder iterierte KL-Transformation genannt, ergibt sich durch folgende Vorgehensweise:

- a) Der hochdimensionale Merkmalvektor eines Musters wird mit Hilfe der KL-Transformation in einen niederdimensionalen Merkmalvektor abgebildet.
- b) Der niederdimensionale Vektor wird durch Hinzufügen aller möglichen Produkte aus zwei seiner Komponenten wieder in einen hochdimensionalen Vektor überführt. Diese Operation wird im folgenden Bildung des "vollständig quadratischen Polynoms" genannt.

- c) Der so entstandene hochdimensionale Vektor wird einer erneuten, dem neuen Problem gerechten KL-Transformation unterworfen.

Man gewinnt durch diese Methode der Merkmalsauswahl im Anwendungsfall die Einbeziehung Glieder höherer Ordnung von Merkmalen in die Prüfgrößen der Klassifikatoren. Aus diesem Grund können dann im Merkmalsraum Trennflächen höherer als zweiter Ordnung auftreten. Die Gruppe der zur Verfügung stehenden Merkmalsauswahlverfahren ist also zu ergänzen durch

- v) Merkmalsauswahl durch iterierte Hauptachsentransformation.

Damit standen für die Durchführung der Experimente zwei Datensätze, drei Klassifikatoren und fünf Merkmalsauswahlverfahren zur Verfügung. Damit ergibt sich, daß ein Experiment mit einer festen Zieldimension aus insgesamt 30 Einzelversuchen besteht, da jedes Auswahlverfahren auf beide Datensätze angewandt werden muß und die so erhaltenen Ergebnisse mit allen drei Klassifikatoren auf ihre Leistungsfähigkeit hin überprüft und verglichen werden müssen.

Eine Sonderstellung nimmt die KL-Transformation bei den Merkmalsauswahlverfahren ein. Dies hat drei Gründe. Zum ersten wurden, wie in Abschnitt 4.2 ausgeführt, beide Datensätze vorab KL-transformiert. Eine ausführliche Begründung wurde angeführt. Damit beeinflußt die KL-Transformation bis zu einem gewissen Grade die Ergebnisse aller anderen Auswahlverfahren. Es zeigt sich aber, daß man aus dieser Tatsache eher Vorteile ziehen kann, als daß man Nachteile erfährt.

Zum zweiten ist die KL-Transformation in der Mustererkennung ein Standardverfahren, das in vielen Fällen zitiert und angewandt wird /N2, N3, S1, F1, O2/. Bereits damit ist sie als Vergleichsverfahren qualifiziert. Es wurde aber auch

gezeigt, daß die Klassifikationsergebnisse mit den Merkmalen, die die KL-Entwicklung liefert, immer besser sind als die Klassifikationsergebnisse, die man erhält, wenn man als Merkmale die Entwicklungskoeffizienten einer problemunabhängigen Reihenentwicklung nimmt /A1/. Mit dem Vergleich eines neuen Verfahrens mit der KL-Entwicklung ist damit also auch ein Vergleich mit einem Standard-Niveau möglich.

Der dritte Grund für die Sonderstellung der KL-Transformation im Rahmen der vorliegenden Arbeit ist, daß aus ihr eine Strategie für die Berechnung der Merkmalsauswahlverfahren ii) bis iv) abgeleitet werden kann. Diese wird in Abschnitt 4.4 noch zu erläutern sein.

4.4 Zur Implementierung der Merkmalsauswahlverfahren

Drei der im Rahmen dieser Arbeit betrachteten Merkmalsauswahlverfahren führen bei der Bestimmung der linearen Auswahlfunktion $\underline{A}$ auf eine Optimierungsaufgabe, die mit numerischen Verfahren gelöst werden muß. Es handelt sich um die Auswahlverfahren ii) (NVK-optimale Merkmale), iii) (MMA-optimale Merkmale) und iv) (DIV-optimale Merkmale). Ehe auf die Besonderheiten der zu diesen Auswahlverfahren gehörigen Zielfunktionen eingegangen wird, sollen die in Frage kommenden numerischen Optimierungsverfahren kurz vorgestellt werden.

4.4.1 Verfahren der numerischen Optimierung

Die betrachtete Grundaufgabe sei

$$F(\underline{A}) \stackrel{!}{=} \text{Min} . \quad (4-3)$$

Dabei sei $\underline{A}$ ein Vektor im $\mathbb{R}^n$. Prinzipiell ist das globale Minimum von F im $\mathbb{R}^n$ zu suchen. Alle hier besprochenen Ver-

fahren finden jedoch immer nur ein lokales Minimum.

Die hinreichenden und notwendigen Bedingungen für das Vorliegen eines lokalen Minimum an der Stelle $\underline{A}_0$ sind /H3/

$$\left. \nabla_{\underline{A}} F(\underline{A}) \right|_{\underline{A} = \underline{A}_0} = \underline{0} \quad (4-4)$$

und

$$\left. \nabla_{\underline{A}} \nabla_{\underline{A}, t} F(\underline{A}) \right|_{\underline{A} = \underline{A}_0} \text{ positiv semidefinit} \quad (4-5)$$

Gleichung (4-4) definiert ein System von n nichtlinearen Gleichungen in $\underline{A}$. Eine Nullstelle dieses Gleichungssystems kann mit Hilfe der Newton-Raphson-Methode bestimmt werden /B2/. Diese Methode verlangt allerdings die Berechnung der zweiten partiellen Ableitungen der Zielfunktion. Sie wurde aus diesem Grund für keines der drei Merkmalsauswahlverfahren verwendet.

Für die Zielfunktionen, die im Rahmen dieser Arbeit zu optimieren sind, kommen als Optimierungsverfahren die sogenannten Abstiegsverfahren in Frage. Diese Verfahren zerfallen in aller Regel in drei Teile, die jeweils gesondert betrachtet werden können: die Bestimmung einer Abstiegsrichtung, die Suche nach einem Extrempunkt der Funktion F entlang der Abstiegsrichtung und die Feststellung der Konvergenz des Verfahrens. Die Arbeitsweise der Verfahren kann für Funktionen im $\mathbb{R}^2$ veranschaulicht werden. Man denkt sich die Funktion $F(a_1, a_2)$ als Geländeprofil dargestellt, das heißt, über jedem Punkt (a_1, a_2) der Ebene wird der Wert $F(a_1, a_2)$ als Höhe dargestellt. Die Optimierungsaufgabe verlangt nun, ausgehend von einem Startpunkt eine "Talsole" in dem so definierten Gebirge zu finden.

Mathematisch gesehen ist die Abstiegsrichtung ein Vektor $\underline{r}$ im $\mathbb{R}^n$. Die Suche nach dem Minimum "entlang $\underline{r}$ " entspricht dem Aufsuchen des Minimums der Funktion

$$F'(t) = F(\underline{A} + t \cdot \underline{x}) \quad (4-6)$$

$F'(t)$ ist eine Funktion von nur noch einer Variablen. Diese Funktionen sind häufig einer analytischen Behandlung zugänglicher als $F(\underline{A})$ selbst /N4/. Ein Abbruchkriterium für Abstiegsverfahren ist zum Beispiel Gleichung (4-4). Diese Gleichung muß bei Vorliegen eines lokalen Extremwertes in jedem Fall erfüllt sein.

Die verschiedenen Abstiegsverfahren unterscheiden sich in der Art, wie die Abstiegsrichtung festgelegt wird. Die Koordinatenabstiegsverfahren lassen als mögliche Abstiegsrichtungen nur Richtungen parallel zu den Basisvektoren des $\mathbb{R}^n$ zu. Die Wahl einer speziellen Richtung, ausgehend von einem aktuellen Standort, bleibt dem Spezialfall überlassen. Das Gradientenabstiegsverfahren bestimmt als Abstiegsrichtung den negativen Gradienten der Zielfunktion. Da der positive Gradient einer Zielfunktion immer die Richtung angibt, entlang der die Funktion am schnellsten wächst, bezeichnet man die Gradientenabstiegsmethode auch als Methode des steilsten Abstiegs. Die Gradientenabstiegsmethode kann als Optimierungsverfahren nur dann eingesetzt werden, wenn der Gradient der Zielfunktion berechnet werden kann.

Eine modifizierte Version des Gradientenabstiegs ist die Methode nach Davidon-Fletcher-Powell (DFP-Algorithmus) /R1, F8, H3/. Diese Methode hat den Vorteil, daß sie für spezielle Klassen von Zielfunktionen besonders schnell eine Lösung findet. Die Methode, mit der diese Beschleunigung erreicht wird, ist, ein Protokoll über die bereits durchlaufenen Abstiegsrichtungen zu führen. Neue Abstiegsrichtungen werden in Abhängigkeit von diesem Protokoll und dem Gradienten der Zielfunktion bestimmt. Mathematisch gesehen ist das Protokoll eine $n \times n$ - Matrix. In dieser Matrix wird Schritt für Schritt eine Approximation der Matrix der zweiten partiellen Ableitungen der Zielfunktion aufgebaut. Das Verfahren basiert auf der abgebrochenen Taylor-Entwicklung. Es konver-

giert schnell für solche Funktionen, deren Taylorentwicklung nach den Gliedern zweiter Ordnung mit vernachlässigbarem Fehler abgebrochen werden kann. Für andere Funktionen muß eine Version des Verfahrens mit Neustart verwendet werden /H3/.

Das DFP-Verfahren verlangt, daß der Gradient der Zielfunktion berechnet werden kann. Der Hauptnachteil des Verfahrens liegt jedoch darin, daß die Hilfsmatrix für das Protokoll mit wachsender Variablenzahl schnell die Speicherkapazität verfügbarer Rechenanlagen übersteigt. Diese Hilfsmatrix ist eine symmetrische Matrix von der Ordnung "Anzahl Variable". Für ein Merkmalsauswahlproblem ergibt sich die Anzahl der Variablen als Produkt

$\text{Zieldimension} \times (\text{Ausgangsdimension} - \text{Zieldimension})$.

Eine symmetrische Matrix der Ordnung n benötigt $(n \cdot (n-1))/2$ Speicherplätze. Damit ergibt sich, daß die Hilfsmatrix für das DFP-Verfahren bei einem Merkmalsauswahlproblem, das eine Dimensionsreduktion von 70 auf 30 Merkmale realisieren soll, bereits ca. 720 K Worte belegen würde. Man hat es hier mit einem typischen Problem der Art Rechenzeit auf Kosten von Speicherplatz zu tun.

Die Abstiegsverfahren führen auf das Problem, das Minimum der Zielfunktion entlang einer Richtung aufzusuchen. Es ist mitunter möglich, Lösungen dieses Problems analytisch zu bestimmen. Wenn dies nicht möglich ist, kann man Einschachtelungsalgorithmen wie die goldene-Schnitt-Regel /R1, H3/ verwenden. Man kann auch versuchen, die Funktion nach Gleichung (4-6) durch ein Polynom in t zu approximieren und die Extremstellen dieses Näherungspolynoms als Lösung ansehen. Für die Bestimmung von Merkmalsauswahlfunktionen erwiesen sich jedoch nur Einschachtelungsalgorithmen /R1/ und heuristische Suchverfahren als akzeptabel (siehe Abschnitt 4.4.3).

Bis hierher wurden Verfahren vorgestellt, die ein lokales Minimum im $\mathbb{R}^n$ aufsuchen, ohne daß die Menge der möglichen Lösungspunkte durch zusätzliche Annahmen eingeschränkt wurde. Im Falle der Bestimmung einer Merkmalsauswahlfunktion muß man jedoch Einschränkungen festlegen. Der Grund hierfür ist, daß es im allgemeinen zu jeder Auswahlfunktion $\underline{A}$ eine Schar von anderen Auswahlfunktionen gibt, auf der die Gütefunktion einen konstanten Wert annimmt. Es wird angestrebt, die Lösung der Optimierungsaufgabe nur auf einer Menge von Repräsentanten dieser Scharen zu suchen. Dies führt auf restringierte Optimierungsaufgaben. Für die lineare Merkmalsauswahl wird gerne die Restriktion

$$\|\underline{A}\| = 1 \quad (4-7)$$

angenommen /F1, F6/. Die resultierende Optimierungsaufgabe mit (4-3) als Zielfunktion und (4-7) als Nebenbedingung wird über die zugehörige Lagrangefunktion

$$\phi(\underline{A}, \delta) = F(\underline{A}) + \delta \cdot \|\underline{A}\| \quad (4-8)$$

gelöst. Numerische Verfahren zur Lösung von (4-8) müssen $\underline{A}$ und δ iterativ bestimmen /F6/. Es zeigt sich jedoch, daß im Falle der Bestimmung einer Merkmalsauswahlfunktion das restringierte Problem auf ein Problem ohne Nebenbedingungen zurückgeführt werden kann /D2/. Genau diese Methode wird für die Durchführung der Versuche im Rahmen der vorliegenden Arbeit verwendet.

4.4.2 Besonderheiten bei der Bestimmung linearer Merkmalsauswahlverfahren

Die Bestimmung linearer Merkmalsauswahlfunktionen zeigt zwei Besonderheiten, die es gestatten, die Lösung der entsprechenden Optimierungsaufgaben effektiver zu gestalten.

Die eine Besonderheit folgt aus Eigenschaft (vii) für Güte-

maße in Abschnitt 2.2.1. Diese Eigenschaft fordert, daß Gütemaße ihren Wert nicht ändern, wenn ein Merkmalsraum einer bijektiven Abbildung unterworfen wird. Für Gütemaße, die diese Eigenschaft erfüllen, ist es möglich, die restringierte Optimierungsaufgabe mit Zielfunktion (4-3) und Nebenbedingung (4-8) in eine unrestringierte Optimierungsaufgabe zu überführen. Es ist dazu nötig, die Scharen von Matrizen $\underline{A}$ (lineare Merkmalsauswahl!) zu bestimmen, auf welchen die Zielfunktion ihren Wert nicht ändert und daran anschließend in diesen Scharen geeignete Repräsentanten zu finden.

Lineare, bijektive Abbildungen des $\mathbb{R}^{n'}$ in sich sind Multiplikationen mit nichtsingulären Matrizen n' -ter Ordnung. Sei $\underline{B}$ solch eine Matrix und $\underline{A}$ eine $n \times n'$ Matrix, die im Sinne eines Auswahlverfahren einen optimalen Merkmalvektor $\underline{c}'$ liefert. Dann erfüllt auch der durch Gleichung (4-9) definierte Merkmalvektor $\underline{c}''$ die Optimalitätseigenschaft.

$$\underline{c}'' = \underline{B} \cdot \underline{c}' = \underline{B} \cdot \underline{A}_t \cdot \underline{c} = \underline{B}^* \underline{c} \quad (4-9)$$

Zu einer Schar von Extraktionsmatrizen werden nun alle die Matrizen $\underline{A}$ zusammengefaßt, die mittels Multiplikation mit einer nichtsingulären $n' \times n'$ Matrix ineinander überführt werden können. Als Repräsentant einer solchen Schar wird die Matrix $\tilde{\underline{A}}$ gewählt, die sich in der Form

$$\tilde{\underline{A}} = \begin{pmatrix} \underline{I}_{n'} \\ \tilde{\underline{A}}_r \end{pmatrix} \quad (4-10)$$

darstellen läßt. $\tilde{\underline{A}}$ ist aus zwei Blöcken aufgebaut. Ein Block wird als Einheitsmatrix der Ordnung n' besetzt, der zweite Block ist eine Restmatrix $\tilde{\underline{A}}_r$ der Ordnung $(n - n') \times n'$.

Eine solche Matrix $\tilde{\underline{A}}$ existiert, wenn die n' Spalten von $\underline{A}$ linear unabhängig sind. Dann bilde man aus n' Zeilen von $\underline{A}$ eine Matrix $\underline{E}$. $\underline{E}$ ist wegen der linearen Unabhängigkeit der Spalten von $\underline{A}$ invertierbar. Setzt man nun in (4-9) die Inverse von $\underline{E}$ für $\underline{B}$ ein, so erhält man für $\underline{B}$ unmittelbar die

Darstellung (4-10):

$$\underline{B} = \underline{E}^{-1} \begin{pmatrix} \underline{E} \\ \underline{F} \end{pmatrix} = \begin{pmatrix} \underline{I}_n' \\ \underline{E}^{-1}\underline{F} \end{pmatrix}$$

Die Forderung nach linearer Unabhängigkeit der Spalten von $\underline{A}$ ist im Zusammenhang mit Merkmalsauswahl wünschenswert, da sie sicherstellt, daß nur linear unabhängige Merkmale in der Klassifikationsstufe verwendet werden.

Der Extremwert der Zielfunktion $F(\underline{A})$ muß nun nur noch in Abhängigkeit von $\tilde{\underline{A}}_r$ gesucht werden /R1, D2/. Dies ist aber ein unrestringiertes Problem. Verwendet man als Lösungsverfahren Koordinatenabstieg, so macht sich zudem die geringere Zahl von unbekannten Variablen bemerkbar.

Eine zweite Besonderheit bei der Bestimmung linearer Merkmalsauswahlfunktionen hat Auswirkungen auf die Festlegung der Reihenfolge der Abstiegsrichtungen beim Koordinatenabstieg. Voraussetzung ist allerdings, daß Kenntnisse über die ungefähre Verteilung der Eingangsmerkmale vorliegen. Diese Kenntnisse lassen sich leicht mit Hilfe der KL-Transformation beschaffen. Aus diesem Grund wurde der KL-Transformation auch eine Sonderstellung unter den Merkmalsauswahlverfahren in Abschnitt 4.3 eingeräumt.

Nach einer KL-Transformation kennt man eine Bewertung von Merkmalen. Diese Bewertung ergibt sich aus den Eigenwerten, die zu den betreffenden KL-Koordinaten gehören. Diese Eigenwerte sind gleichzeitig die Streuung der zugehörigen Merkmale. Aus der Ordnung der KL-Koeffizienten nach abnehmender Streuung folgt eine Strategie für das Koordinatenabstiegsverfahren: Man faßt Koordinatenabstieg als schrittweise Ermittlung der Transformationsmatrix $\underline{A}$ bzw. des Teils $\tilde{\underline{A}}_r$ auf. Abstieg in einer bestimmten Variablen bedeutet das Testen des Einflusses einer bestimmten Eingangsgröße auf die Zielfunktion. Sind die Eingangsgrößen KL-Koeffizienten, so wird

ihr erwarteter Einfluß umso höher eingeschätzt, je größer der zugehörige Eigenwert ist.

Hieraus ergibt sich eine einfache Strategie für die Durchführung des Koordinatenabstiegsverfahrens. Auswahl einer Koordinate und Abstieg entlang einer solchen Koordinate bedeutet nichts anderes als Auswahl eines Elements der Matrix $\tilde{A}_r$ und die Auswertung der Zielfunktion für geeignete Werte dieses Matrixelements. Die Matrixelemente der i-ten Zeile von $\tilde{A}$ bestimmen aber den Einfluß der i-ten Eingangsvariablen auf die Ausgangsmerkmale c'_i der Extraktionsfunktion, die ja durch

$$\underline{c}' = \underline{A}_t \cdot \underline{c}$$

gegeben ist. Eine einfache Strategie ist, zuerst diejenigen Matrixelemente zu variieren, die solche Eingangsgrößen einbeziehen, deren erwarteter Einfluß am höchsten eingeschätzt wird. Hieraus ergibt sich zusammen mit der oben dargelegten Bewertung der Eingangsgrößen für KL-Koeffizienten, daß $\tilde{A}_r$ sinnvollerweise zeilenweise von oben nach unten abgearbeitet wird.

Ein besonderer Vorteil dieses Verfahrens ist, daß es nach der Abarbeitung der Zeile i eine Zwischenlösung liefert, die als optimale Abbildung vom $\mathbb{R}^i$ in den $\mathbb{R}^{n'}$ angesehen werden kann. Es ist somit möglich, ausgehend von der Startmatrix

$$\underline{A}_s = \begin{pmatrix} \underline{I}_{n'} \\ \underline{0} \end{pmatrix}$$

schrittweise eine Folge von linearen Transformationen zu ermitteln, die immer mehr die Eingangsgrößen berücksichtigen.

4.4.3 Realisierung der Auswahlverfahren

Auf dem Hintergrund von Abschnitt 4.4.2 wird im folgenden

auf die Realisierung der Auswahlverfahren (ii) bis iv) aus Abschnitt 4.3 eingegangen.

4.4.3.1 DIV-optimale Merkmale

Die Definitionsgleichung für die Divergenz zwischen zwei Verteilungen kann Tabelle 2-1 entnommen werden. Setzt man in diese Formel die klassenbedingten Dichten $w(\underline{c}|\Omega_\lambda)$ ein und nimmt weiterhin an, daß diese Dichten normal sind, so kann die Integration analytisch ausgeführt werden und man erhält /F1, R1/

$$D_{\kappa\lambda} = \frac{1}{2}(\underline{\mu}_\kappa - \underline{\mu}_\lambda)'_t (\underline{K}_\kappa^{-1} + \underline{K}_\lambda^{-1}) (\underline{\mu}_\kappa - \underline{\mu}_\lambda) + \frac{1}{2} \text{Sp}(\underline{K}_\kappa^{-1} \underline{K}_\lambda + \underline{K}_\lambda^{-1} \underline{K}_\kappa - 2\underline{I}) \quad (4-11)$$

Aus den transformierten Merkmalvektoren

$$\underline{c}' = \underline{A}_t \underline{c} \quad (4-12)$$

erhält man die transformierten Mittelwertvektoren und Kovarianzmatrizen

$$\underline{\mu}_\kappa' = \underline{A}_t \underline{\mu}_\kappa \quad (4-13)$$

$$\underline{K}_\kappa' = \underline{A}_t \underline{K}_\kappa \underline{A} \quad (4-14)$$

Damit ergibt sich als zu maximierende Funktion

$$D_{\kappa\lambda}(\underline{A}) = \frac{1}{2}(\underline{\mu}_\kappa - \underline{\mu}_\lambda)'_t \underline{A} \{ \underline{A}_t \underline{K}_\kappa \underline{A} \}^{-1} + (\underline{A}_t \underline{K}_\lambda \underline{A})^{-1} \}_t \underline{A} (\underline{\mu}_\kappa - \underline{\mu}_\lambda) + \frac{1}{2} \text{Sp} \{ (\underline{A}_t \underline{K}_\kappa \underline{A})^{-1} (\underline{A}_t \underline{K}_\lambda \underline{A}) + (\underline{A}_t \underline{K}_\lambda \underline{A})^{-1} (\underline{A}_t \underline{K}_\kappa \underline{A}) - 2\underline{I} \} \quad (4-15)$$

Die Erweiterung auf k Klassen liefert die mittlere Divergenz über alle Klassenpaare.

$$D(\underline{A}) = \frac{1}{k \cdot (k-1)} \sum_{\kappa=1}^k \sum_{\lambda=1}^k D_{\kappa\lambda}(\underline{A}) \quad (4-16)$$

Es ist gezeigt worden, daß $D(\underline{A})$ in die spezielle Form

$$D(\underline{A}) = \frac{1}{k \cdot (k-1)} \text{Sp} \left\{ \sum_{\kappa=1}^1 (\underline{A}_{t-\kappa} \underline{A})^{-1} (\underline{A}_{t-\kappa} \underline{A}) \right\} - n' \quad (4-17)$$

mit

$$\underline{S}_{\kappa} = \sum_{\substack{\lambda=1 \\ \lambda \neq \kappa}}^k \left[\frac{\underline{K}_{\lambda}}{\lambda} + (\underline{\mu}_{\kappa} - \underline{\mu}_{\lambda}) (\underline{\mu}_{\kappa} - \underline{\mu}_{\lambda}) \underline{t} \right] \quad (4-18)$$

überführt werden kann /R1, D2/. Aus (4-17) kann eine Gleichung für den Gradienten von $D(\underline{A})$ abgeleitet werden /R1/.

$$\begin{aligned} \nabla_{\underline{A}} D(\underline{A}) &= \frac{2}{k \cdot (k-1)} \\ &\cdot \sum_{\kappa=1}^k \left[\underline{S}_{\kappa} \underline{A} - \underline{K}_{\kappa} \underline{A} (\underline{A}_{t-\kappa} \underline{A})^{-1} (\underline{A}_{t-\kappa} \underline{A}) \right] \cdot (\underline{A}_{t-\kappa} \underline{A})^{-1} \end{aligned} \quad (4-19)$$

Es ist deshalb möglich, solche Abstiegsverfahren zur Ermittlung der optimalen Transformationsmatrix $\underline{A}$ einzusetzen, die die Berechnung des Gradienten verlangen. Realisiert wurden der DFP-Algorithmus /R1/ und der Gradientenabstieg. Der Gradientenabstieg wird für Reduktionen mit hoher Variablenzahl eingesetzt, da in diesen Fällen der Speicherplatzbedarf des DFP-Algorithmus zu hoch für die zur Verfügung stehende Maschine ist. In jedem Fall wird die Extremwertsuche entlang einer Abstiegsrichtung mit Hilfe der goldenen-Schnitt-Regel durchgeführt /R1, H3/.

Aus Formel (4-11) und aus Formel (4-17) kann abgeleitet werden, daß die Divergenz invariant gegenüber nichtsingulären, linearen Transformationen ist. Die Optimierung von (4-16) kann deshalb gemäß Abschnitt 4.4.2 unrestringiert durchgeführt werden. Als Startpunkt wurde in allen Fällen die Nullmatrix für $\tilde{\underline{A}}_r$ aus (4-10) gewählt. Dies bedeutet nichts anderes, als daß eine Projektion auf die ersten n' Komponenten des Urbildraumes gewählt wird. Da im Rahmen

dieser Arbeit der Urbildraum aus dem Raum der ersten n KL-Koeffizienten besteht, heißt dies, daß die "Startabbildung" die Abbildung auf die ersten n' KL-Koeffizienten ist.

4.4.3.2 MMA-optimale Merkmale

Die Schlüsselgrößen für die klassifikatorbezogene Merkmalsauswahl beim MMA sind die R_k nach Gleichung (3-17) bzw. die $R_{k\lambda}$ nach Gleichung (3-15). Eine naheliegende Idee ist, die R_k oder die $R_{k\lambda}$ oder einen Mittelwert davon zu maximieren. Da aber durch Gleichung (3-25) ein direkter Bezug zur Fehlerwahrscheinlichkeit des MMA hergestellt werden kann, wurde Gleichung (3-25) als Zielfunktion gewählt. Damit erhält man

$$G(\underline{A}) = \sum_{k=1}^K P(\Omega_k) \frac{n'}{R_k(\underline{A})} \quad (4-20)$$

Die a priori Wahrscheinlichkeiten wurden in den durchgeführten Versuchen zu

$$P(\Omega_k) = \frac{1}{K} \quad (4-21)$$

angenommen. $R_k(\underline{A})$ ergibt sich aus der Definition von R_k und Gleichung (3-28).

Gleichung (3-28) zeigt, daß R_k gegenüber linearen, nicht-singulären Abbildungen des $\mathbb{R}^n$ invariant ist. Man braucht dazu in Gleichung (3-28) $\underline{A}$ nur als nichtsinguläre Matrix der Ordnung n einsetzen. Damit ist auch (4-20) gegen eine nichtsinguläre Abbildung invariant. Es ist somit möglich, die Optimierung von $G(\underline{A})$ nach (4-20) unrestringiert durchzuführen. Als Startpunkt wird wie bei der Bestimmung der DIV-optimalen Merkmale die Nullmatrix für $\underline{A}_r$ gewählt.

Als Abstiegsverfahren zur Berechnung der MMA-optimalen Merkmale wurde der Koordinatenabstieg gewählt. Der Grund für

diese Wahl ist, daß der Gradient von $G(\underline{A})$ nach Gleichung (4-20) nicht zur Verfügung steht. Die Ableitung der Funktion

$$\tilde{G}(t) = G(\underline{A} + t \cdot \underline{r}) \quad (4-22)$$

kann bei festem $\underline{r}$ durch die Differenzenformel

$$\frac{\tilde{G}(t) - \tilde{G}(t + h)}{h} \quad (4-23)$$

approximiert werden. Mit Hilfe der Formel (4-23) ist es damit möglich, nach Auswahl einer Koordinate die Abstiegsrichtung entlang dieser Koordinate ohne Berechnung des Gradienten der Zielfunktion festzulegen. Die Auswertung der Funktion $G(\underline{A})$ erfordert die Berechnung der Größen $R_{\kappa\lambda}^2$. Die Berechnung dieser Größen stellt sich bereits als Optimierungsaufgabe dar. Diese Aufgabe lautet gemäß Gleichung (3-16) aus Abschnitt 3.3

$$f(\underline{c}) = (\underline{c} - \underline{\mu}_{\kappa})^T K_{\kappa}^{-1} (\underline{c} - \underline{\mu}_{\kappa}) \stackrel{!}{=} \text{Min}$$

$$g(\underline{c}) = (\underline{c} - \underline{\mu}_{\kappa})^T K_{\kappa}^{-1} (\underline{c} - \underline{\mu}_{\kappa}) - (\underline{c} - \underline{\mu}_{\lambda})^T K_{\lambda}^{-1} (\underline{c} - \underline{\mu}_{\lambda}) = 0 \quad (4-24)$$

Aus der Optimierungstheorie ist bekannt, daß solche nicht-lineare Aufgaben über die zugehörige Lagrangefunktion

$$\phi(\underline{c}, \delta) = f(\underline{c}) + \delta \cdot g(\underline{c}) \quad (4-25)$$

gelöst werden können. Im Rahmen der vorliegenden Arbeit wurde jedoch ein Verfahren bevorzugt, das sich aus der Methode der projizierten Gradienten /H3/ ableiten läßt. Das Verfahren ist insbesondere geometrisch anschaulich.

Das Grundproblem des Verfahrens ist, ausgehend von einem Punkt $\underline{c}_0$ auf der Klassengrenze einen Punkt $\underline{c}_1$ aufzusuchen, der ebenfalls auf der Klassengrenze liegt (d. h., der Restriktion in Gleichung (4-24) genügt) und der näher bei $\underline{\mu}_{\kappa}$ liegt. Diese Aufgabe wird in folgende Teilschritte zerlegt:

- a) Bestimme einen Punkt $\underline{c}_0$, der in derjenigen Hyperebene liegt, die die Klassengrenze in $\underline{c}_0$ berührt und der näher bei $\underline{\mu}_k$ liegt.
- b) Suche zur Lösung $\underline{c}_0$ von Teilschritt a) einen Punkt $\underline{c}_1$ auf der Klassengrenze

Teilschritt a) wird mit Hilfe der Methode des projizierten Gradienten gelöst. In Anlehnung an diese Methode wird ein Lösungspunkt $\underline{c}_0$ nur auf derjenigen Geraden gesucht, die durch die Projektion des Gradienten der Zielfunktion auf die in a) genannte Hyperebene festgelegt wird. Für die Aufgabe (4-24) kann $\underline{c}_0$ analytisch bestimmt werden.

Hierzu wird zunächst die Projektion des Gradienten der Zielfunktion $f(\underline{c})$ auf die erwähnte Hyperebene benötigt. Es ist

$$\nabla_{\underline{c}} f(\underline{c}_0) = 2\underline{K}_k^{-1}(\underline{c}_0 - \underline{\mu}_k) \quad (4-26)$$

der Gradient der Zielfunktion im Punkt $\underline{c}_0$. Weiter ist

$$\underline{n} = 2\underline{K}_k^{-1}(\underline{c}_0 - \underline{\mu}_k) - 2\underline{K}_\lambda^{-1}(\underline{c}_0 - \underline{\mu}_\lambda) \quad (4-27)$$

der Normalenvektor der gesuchten Hyperebene. $\underline{n}$ ist der Gradient der Restriktion aus (4-24).

Ist $\underline{n}$ der Normalenvektor einer Ebene, so ist die Matrix

$$\underline{P} = (\underline{I} - \frac{\underline{n}\underline{n}^t}{\underline{n}^t\underline{n}}) \quad (4-28)$$

der Projektionsoperator auf dieser Ebene, d. h., die Multiplikation eines Vektors $\underline{c}$ mit $\underline{P}$ liefert gerade die Projektion von $\underline{c}$ auf die Ebene mit $\underline{n}$ als Normalenvektor. Damit ist

$$\underline{r} = \underline{P} \cdot \nabla_{\underline{c}} f(\underline{c}_0) \quad (4-29)$$

die gesuchte Projektion des Gradienten der Zielfunktion auf die Tangentialhyperebene aus Schritt a).

Gesucht ist nun das Minimum der Funktion

$$l(t) = f(\underline{c}_0 + t \cdot \underline{r}) . \quad (4-30)$$

Dieses wird durch Auflösen der Gleichung

$$\frac{dl(t)}{dt} = 0 \quad (4-31)$$

nach t bestimmt. (4-31) liefert die Gleichung

$$2\underline{r}_t \underline{K}_k^{-1} (\underline{c}_0 - \underline{\mu}_k) - 2t \underline{r}_t \underline{K}_k^{-1} \underline{r} = 0 . \quad (4-32)$$

Die Auflösung ergibt

$$t_0 = \frac{\underline{r}_t \underline{K}_k^{-1} (\underline{c}_0 - \underline{\mu}_k)}{\underline{r}_t \underline{K}_k^{-1} \underline{r}} , \quad (4-33)$$

der Lösungspunkt $\underline{c}_0$ ist somit eindeutig.

$$\underline{c}_0 = \underline{c}_0 + t_0 \cdot \underline{r} \quad (4-34)$$

Teilschritt a) ist damit gelöst. Es bleibt nun noch die Aufgabe, mit Hilfe von $\underline{c}_0$ einen Punkt $\underline{c}_1$ auf der Klassengrenze zu finden. Als Lösung dieser Aufgabe wird hier derjenige Punkt $\underline{c}_1$ gewählt, der sich als Schnittpunkt der Geraden, die durch $\underline{\mu}_k$ und $\underline{c}_0$ festgelegt ist, mit der Klassengrenze ergibt. Diese Gerade hat die Gleichung

$$\underline{c}(t) = \underline{\mu}_k + t \cdot (\underline{c}_0 - \underline{\mu}_k) . \quad (4-35)$$

$\underline{c}(t)$ eingesetzt in die Restriktionsgleichung $g(\underline{c}) = 0$ aus (4-24) liefert eine in t quadratische Gleichung:

$$\begin{aligned} & t^2 \cdot \left[(\underline{c}_0 - \underline{\mu}_k)_t (\underline{K}_k^{-1} - \underline{K}_\lambda^{-1}) (\underline{c}_0 - \underline{\mu}_k) \right] \\ & - 2t \cdot \left[(\underline{\mu}_k - \underline{\mu}_\lambda)_t \underline{K}_\lambda^{-1} (\underline{c}_0 - \underline{\mu}_k) \right] \\ & - (\underline{\mu}_k - \underline{\mu}_\lambda)_t \underline{K}_\lambda^{-1} (\underline{\mu}_k - \underline{\mu}_\lambda) = 0 \end{aligned} \quad (4-36)$$

Die Auflösung dieser Gleichung bedeutet keine Schwierigkeit. Es kann jedoch vorkommen, daß keine reelle Lösung existiert. In diesem Fall schneidet die Gerade nach (4-35) die Klassengrenze nicht. Ein solcher Fall ist in Bild 4-1 mit der Geraden g_1 gegeben. $\underline{c}_0$ ist dort der Punkt auf der "Hyperebene" h mit kleinstem Abstand zum Mittelvektor $\underline{\mu}_1$ der Klasse Ω_1 . Die Gerade durch $\underline{\mu}_1$ und $\underline{c}_0$ schneidet die Klassengrenze e (im gezeigten Fall eine Ellipse) nicht.

Prinzipiell kann in solchen Fällen kein Lösungspunkt $\underline{c}_1$ ermittelt werden. Dies ist jedoch im Zusammenhang mit dem übergeordneten Verfahren zur Lösung von (4-24) ohne Bedeutung. Dieses übergeordnete Verfahren nutzt aus, daß die beiden Optimierungsaufgaben

$$f_K(\underline{c}) = (\underline{c} - \underline{\mu}_K)_t K_K^{-1} (\underline{c} - \underline{\mu}_K) \stackrel{!}{=} \text{Min} \quad (4-37)$$

und

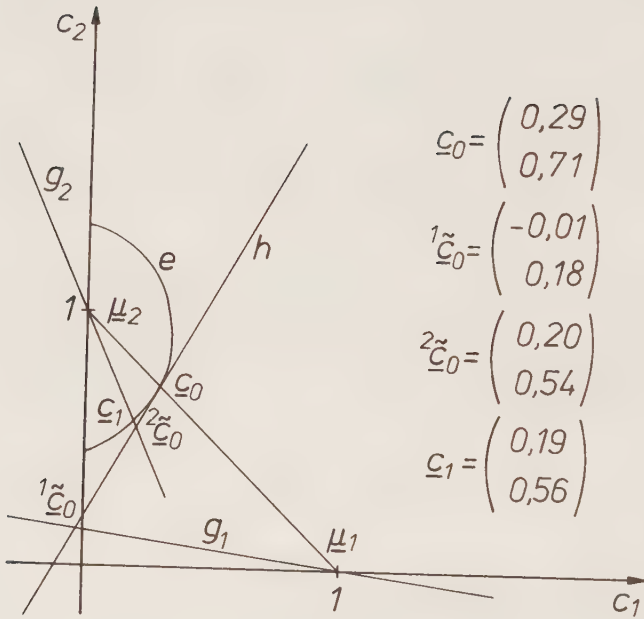
$$f_\lambda(\underline{c}) = (\underline{c} - \underline{\mu}_\lambda)_t K_\lambda^{-1} (\underline{c} - \underline{\mu}_\lambda) \stackrel{!}{=} \text{Min} \quad (4-38)$$

unter der gemeinsamen Restriktion

$$f_K(\underline{c}) - f_\lambda(\underline{c}) = 0 \quad (4-39)$$

die gleiche Lösung besitzen. Dies bedeutet, daß der "Abstieg" bezüglich (4-37) auf der Klassengrenze auch ein "Abstieg" bezüglich (4-38) ist. Prinzipiell kann man also die gemeinsame Lösung durch Abstieg bezüglich (4-37) oder durch Abstieg bezüglich (4-38) suchen.

Das im Rahmen dieser Arbeit benutzte Verfahren führt ausgehend von einem Punkt $\underline{c}_0$ auf der Klassengrenze beide möglichen Abstiegschritte durch. Auf diese Weise erhält man zwei Kandidaten für $\underline{c}_1$. Im Bild 4-1 ist eine entsprechende Konfiguration im $\mathbb{R}^2$ dargestellt. Als endgültige Lösung $\underline{c}_1$ wird der Kandidat gewählt, der den kleineren Wert für die Zielfunktion ergibt. Wie oben dargelegt, kann es vorkommen,



$$\Omega_1: \quad \underline{\mu}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}; \quad \underline{K}_1 = \begin{pmatrix} 10 & 0 \\ 0 & 3 \end{pmatrix}$$

$$\Omega_2: \quad \underline{\mu}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}; \quad \underline{K}_2 = \begin{pmatrix} 0,5 & 0 \\ 0 & 2 \end{pmatrix}$$

e: Klassengrenze

Bild 4-1 Abstieg auf einer Klassengrenze

daß einer der Kandidaten nicht ermittelt werden kann, weil Gleichung (4-36) keine Lösung besitzt. In diesem Fall wird der einzige zur Verfügung stehende Kandidat als Lösung genommen.

Es kann nicht vorkommen, daß keiner der beiden Abstiegschritte eine Lösung liefert. Dies sieht man daraus, daß

- i) $\underline{\mu}_K$ und $\underline{\mu}_\lambda$ auf verschiedenen "Seiten" der Klassengrenze $\underline{c}_0$ liegen
- und
- ii) die Hyperebene gemäß Gleichung (4-27) die Klassengrenze nicht schneidet und somit alle Kandidaten $\underline{c}_0$ auf der gleichen Seite der Hyperebene liegen.

Einer der Klassenmittelwertvektoren $\underline{\mu}_K$ und $\underline{\mu}_\lambda$ muß somit aus der Sicht der $\underline{c}_0$ auf der anderen Seite der Klassengrenze liegen. Die Verbindungsgerade zu diesem der beiden Punkte muß deshalb die Klassengrenze schneiden.

Mit Hilfe der besprochenen Berechnung der Größen $R_{K\lambda}^2$ wird nun $G(\underline{A})$ gemäß Gleichung (4-20) nach folgendem Verfahren berechnet:

- a) berechne die transformierten Mittelwertvektoren und Kovarianzmatrizen
- b) berechne mit Hilfe der Ergebnisse aus a) alle $R_{K\lambda}^2(\underline{A})$ mit dem oben beschriebenen Verfahren
- c) berechne $R_K^2(\underline{A})$ als Minimum der Zahlen $\{R_{K\lambda}^2(\underline{A}) \mid \lambda = 1, \dots, k\}$
- d) berechne $G(\underline{A})$ nach Formel (4-20)

Der Koordinatenabstiegsalgorithmus zur Minimierung von $G(\underline{A})$ wird nun wie folgt gesteuert: Die Reihenfolge der Abstiegskoordinaten ergibt sich aus der Anordnung der beeinflussten Variablen in der Transformationsmatrix. Nach der Festlegung einer Abstiegsordinate muß die zugehörige Variable in der

Transformationsmatrix $\underline{A}$ variiert werden. Das Ziel ist, den Wert für diese Variable zu finden, der $G(\underline{A})$ minimiert.

Dieses Problem kann weder analytisch noch numerisch exakt gelöst werden. Daher wurde ein heuristisches Verfahren angewandt, das $G(\underline{A})$ für einige, systematisch ausgewählte Matrizen $\underline{A}_m$ berechnet und als beste Matrix $\underline{A}_0$ diejenige festlegt, die den kleinsten Wert für $G(\underline{A})$ liefert. Die Matrizen $\underline{A}_m$ erhält man, wenn man das Element a_{ij} , welches durch das Koordinatenabstiegsverfahren als Abstiegsordinate festgelegt wurde, durch die Zahlen

$$a_m = a_{ij} + m \cdot h, \quad m = 0, \pm 1, \pm 2, \dots \\ h > 0 \quad (4-40)$$

ersetzt. a_{ij} repräsentiert hier den alten Wert der zugehörigen Komponente.

Im Rahmen der Versuche zur vorliegenden Arbeit wurde folgende Strategie verwendet:

1. Es werden $G(\underline{A}_0)$, $G(\underline{A}_1)$, $G(\underline{A}_{-1})$ berechnet.
2. Ist $G(\underline{A}_0) < G(\underline{A}_1)$ und gleichzeitig $G(\underline{A}_0) < G(\underline{A}_{-1})$, so wird $\underline{A}_0$ als Lösung angesehen.
3. Ist die Bedingung in 2. nicht erfüllt, so werden (L-1) weitere Werte $G(\underline{A}_m)$ berechnet, und zwar

$$G(\underline{A}_m), \quad m = 2, 3, \dots, L \\ \text{falls } G(\underline{A}_1) < G(\underline{A}_{-1}) \\ \text{andernfalls} \\ G(\underline{A}_m), \quad m = -2, -3, \dots, -L$$

Unter den so berechneten L Zahlen wird die kleinste herausgesucht und die zugehörige Matrix $\underline{A}_m$ als Lösung der Minimumsuche entlang einer Koordinatenachse eingesetzt.

Die so festgelegte Methode erweist sich bei der Durchführung der Experimente als effektiv. L und h sind Parameter des Verfahrens und für die numerische Durchführung von großer

Bedeutung. Die benötigte Rechenzeit ist weitgehend linear in L, Abweichungen vom linearen Zusammenhang ergeben sich als Folge der Bedingung unter Punkt 2. L und h zusammen bestimmen die Genauigkeit des Algorithmus bei der Ermittlung des Minimums entlang einer Abstiegsrichtung.

4.4.3.3 NVK-optimale Merkmale

Die Zielfunktion zur Ermittlung der NVK-optimalen Merkmale erhält man, wenn man in Gleichung (3-4) und Gleichung (3-5) für $w(\underline{c}|\underline{\Omega}_k)$ die Normalverteilungsdichte einsetzt und wieder lineare Merkmalsauswahl annimmt. Die so entstehende Zielfunktion enthält eine mehrdimensionale Integration, die wegen der nicht einfach ausdrückbaren Integrationsgrenzen nicht analytisch ausgewertet werden kann.

$$V(\underline{A}) = \sum_{\lambda=1}^k \int_{\Omega_{\lambda}^0(\underline{A})} u_{\lambda}(\underline{c}, \underline{A}) d\underline{c}$$

$$u_{\lambda}(\underline{c}, \underline{A}) = \sum_{\substack{\kappa=1 \\ \kappa \neq \lambda}}^k p(\underline{\Omega}_{\kappa}) w(\underline{c}|\underline{\Omega}_{\kappa}, \underline{A})$$

$$w(\underline{c}|\underline{\Omega}_{\kappa}, \underline{A}) = N(\underline{A}_{-t} \underline{\mu}_{\kappa}, \underline{A}_{-t} \underline{K}_{\kappa} \underline{A}_{-t}) \quad (4-41)$$

Dabei bezeichnet $N(\underline{\mu}, \underline{K})$ die Normalverteilung mit Mittelwertvektor $\underline{\mu}$ und Kovarianzmatrix $\underline{K}$.

Die Zielfunktion $V(\underline{A})$ in (4-41) hat folgende Eigenschaften:

- Sie kann wegen der auftretenden mehrdimensionalen Integration nicht analytisch ausgewertet werden
- In /F5/ wurde gezeigt, daß der Gradient $\nabla_{\underline{A}} V(\underline{A})$ in geschlossener Form dargestellt werden kann.
- $V(\underline{A})$ ist die Bayesfehlerrate nach Ausführen der durch $\underline{A}$ definierten linearen Transformation. Diese Fehlerrate ist unabhängig von nichtsingulären, linearen Transformationen des zugehörigen Merkmalraumes.

Eigenschaft a) hat zur Folge, daß $V(\underline{A})$ numerisch bestimmt werden muß. Aber auch die numerische Behandlung von $V(\underline{A})$ ist sehr aufwendig. Im Rahmen der vorliegenden Arbeit wurde daher vorgezogen, den Wert $V(\underline{A})$ mit Hilfe einer Lernstichprobe zu schätzen. $V(\underline{A})$ ist im vorliegenden Fall die Fehlerrate des verwendeten Klassifikators. Diese kann aber mit Hilfe der Lernstichprobe geschätzt werden.

Eigenschaft b) legt die Verwendung eines Gradientenabstiegsverfahrens nahe. Prinzipiell ist dies möglich; in /F6/ wird über eine Realisierung berichtet. Grundsätzlich erhält die Formel für den Gradienten von $V(\underline{A})$ eine mehrdimensionale Integration, die wieder auf eine Schätzung mit Hilfe einer Lernstichprobe zurückgeführt werden muß. Das prinzipielle Problem bei der praktischen Realisierung des Verfahrens mit Gradientenabstieg liegt aber gerade in der Schätzung des Gradienten. Dieses Problem hat seine Ursache darin, daß zur Schätzung des Gradienten nur diejenigen Muster aus der Lernstichprobe beitragen, die nach der linearen Abbildung falsch klassifiziert werden. Bei kleinen Fehlerwahrscheinlichkeiten kann deshalb trotz umfangreicher Lernstichprobe der Fall eintreten, daß der Gradient von $V(\underline{A})$ mit wenigen Mustern geschätzt werden muß. Im Rahmen der vorliegenden Arbeit wurde deshalb auf die Berechnung des Gradienten verzichtet. Stattdessen wurde als Lösungsverfahren das Koordinatenabstiegsverfahren gewählt. Realisiert wurde das Verfahren ganz analog der Beschreibung, die in Abschnitt 4.4.3.2 gegeben wurde.

Bedingung c) gestattet gemäß Abschnitt 4.4.2, die Lösung über eine unrestringierte Optimierungsaufgabe zu suchen. Die Gültigkeit der Bedingung c) sieht man an den Prüfgrößen des NVK gemäß Gleichung (2-13). Ersetzt man dort den Merkmalvektor $\underline{c}$ durch

$$\underline{c}' = \underline{B}\underline{c} \quad , \quad (4-42)$$

wobei $\underline{B}$ eine nichtsinguläre, quadratische Matrix ist, so

wird aus den Prüfgrößen

$$\begin{aligned} u'_k(\underline{B}\underline{c}) &= -\frac{1}{2}(\underline{c} - \underline{\mu}_k)' \underline{B}_t \underline{B}_t^{-1} \underline{K}_k \underline{B}_t^{-1} \underline{B}(\underline{c} - \underline{\mu}_k) + \gamma_k - \frac{1}{2} \ln |\underline{B}|^2 \\ &= u'_k(c) - \frac{1}{2} \ln |\underline{B}|^2 \end{aligned} \quad (4-43)$$

Der Term $\frac{1}{2} \ln |\underline{B}|^2$ beeinflußt die Entscheidung des Klassifikators nicht. $V(\underline{A})$ ist damit invariant gegen nichtsinguläre, lineare Transformationen.

4.5 Einige Resultate

Die experimentellen Untersuchungen wurden an der TR440 des Regionalen Rechenzentrums Erlangen durchgeführt. Hauptschwierigkeit bei der Durchführung der Untersuchungen war der hohe Bedarf an Rechnerkernzeit für die Berechnung der Transformationsmatrizen für die lineare Merkmalsauswahl bei MMA-optimalen und NVK-optimalen Merkmalen.

Der Rechenzeitbedarf für die Ermittlung dieser Matrizen kann wie folgt abgeschätzt werden. Nach dem in Abschnitt 4.4.3.2 beschriebenen Algorithmus sind für jede der $n' \cdot (n - n')$ zu bestimmenden Variablen L Funktionsauswertungen - das sind Berechnungen der Gütemaße - durchzuführen. Es zeigt sich, daß die Gesamtrechenzeit sehr genau durch das Produkt aus Rechenzeit für eine Funktionsauswertung und Anzahl Funktionsauswertungen bestimmt ist. Die für eine Funktionsauswertung benötigte CPU-Zeit wurde bei NVK-optimalen und bei MMA-optimalen Merkmalen für verschiedene n und n' gemessen. Dabei wurde zur Berechnung der NVK-optimalen Merkmale eine reduzierte Lernstichprobe aus 1000 Mustern zugrunde gelegt. Die Rechenzeiten sind in Tabellen 4-1 und 4-2 angegeben. Die Tabellen 4-3 und 4-4 zeigen die geschätzten Rechenzeiten zur Bestimmung der gesamten Transformationsmatrix, wenn man im beschriebenen Algorithmus $L = 10$ (d. h., 10 Funktionsaus-

wertungen pro zu bestimmenden Koeffizienten) annimmt. Der Wert $L = 10$ wurde im Rahmen der vorliegenden Arbeit aus zwei Gründen gewählt:

1. Die notwendigen Rechenzeiten gemäß den Tabellen 4-1 bis 4-4 konnten mit dem verfügbaren Rechner aufgebracht werden
2. $L = 10$ wurde im beschriebenen Algorithmus in Verbindung mit $h = 0,1$ verwendet. Dies bedeutet, daß die Elemente der Transformationsmatrix Werte zwischen -1 und $+1$ annehmen können; die Werte selbst sind Vielfache von $0,1$. Diese Approximation erwies sich als ausreichend. Ein Beleg dafür ist auch das gute Abschneiden der MMA-optimalen Merkmale.

$n \setminus n'$	5	10	15	20
30	12	27	55	85
70	15	37	66	99

Funktionsauswertung, MMA-optimale
Merkmale, Rechenzeit in Sekunden

Tabelle 4-1

$n \setminus n'$	5	10	15	20
30	36	59	82	117
70	46	72	109	147

Funktionsauswertung, NVK-optimale
Merkmale, Rechenzeit in Sekunden

Tabelle 4-2

$n \backslash n'$	5	10	15	20
30	4,2	15	35	47
70	13,5	61,7	151	275

CPU-Zeit für die Berechnung der Transformationsmatrix in Stunden, MMA-optimale Merkmale

Tabelle 4-3

$n \backslash n'$	5	10	15	20
30	12,5	33	51	65
70	42	120	250	410

CPU-Zeit für die Berechnung der Transformationsmatrix in Stunden, NVK-optimale Merkmale

Tabelle 4-4

Um einen möglichst umfassenden Vergleich der fünf Merkmalsauswahlverfahren aus Abschnitt 4.3 durchführen zu können, wurde n und n' so gewählt, daß die Gesamtrechenzeit im Rahmen dessen blieb, was mit dem verfügbaren Rechner durchgeführt werden kann. Als Ausgangsdimension wurde $n = 30$ gewählt. Es wurden also lineare Transformationen gesucht, die den Raum der ersten 30 KL-Koeffizienten (siehe Abschnitt 4.2) in einem Raum niedrigerer Dimension abbilden. Als Zieldimension n' wurden 2, 5, 10, 15 und 20 genommen. Wegen des enormen Rechenzeitbedarfs wurde auf die Bestimmung NVK-optimaler Merkmale mit Zieldimensionen größer als 5 verzichtet. Weiterhin wurde für $n' = 2$ die Eingangsdimension n in einer weiteren Versuchsreihe auf 70 erweitert. Hiermit wurden alle nach Abschnitt 4.2 zur Verfügung stehenden Eingangsgrößen

berücksichtigt.

Anhang A gibt über das Verhalten der Gütefunktionen während der Durchführung der numerischen Optimierung Aufschluß. Es sind die reinen Zahlenwerte der Gütefunktionen jeweils für die ersten KL-Koeffizienten und für die nach dem entsprechenden Kriterium optimalen Merkmale angegeben.

Von weitaus größerem Interesse als die Zahlenwerte der Gütefunktionen sind die Fehlerraten, welche die drei zur Verfügung stehenden Klassifikatoren mit den unterschiedlichen Merkmalen liefern. Die Anhänge B, C und D enthalten das reine Zahlenmaterial. Bei allen Versuchen wurde die Fehlerrate gemessen, die man erhält, wenn Rückweisungen ausgeschlossen sind.

Anhang B enthält die Klassifikationsergebnisse, wenn man die KL-Koeffizienten als Merkmale heranzieht. Sowohl für die Sprachdaten wie für die Schriftzeichen ist das Phänomen erkennbar, daß die Klassifikationsleistung wieder abnimmt, wenn zu viele Merkmale verwendet werden. Die Sprachdaten werden optimal mit 45 KL-Koeffizienten bei einer Fehlerwahrscheinlichkeit von 0,99 % klassifiziert, die Schriftzeichen mit 35 Merkmalen bei einer Fehlerrate von 1,20 %. In beiden Fällen liefert der Normalverteilungsklassifikator diese optimalen Klassifikationsleistungen.

Aus den Tabellen in Anhang B ist weiter ersichtlich, daß für das vorliegende Datenmaterial der Normalverteilungsklassifikator sowohl dem modifizierten Minimum-Abstand-Klassifikator als auch dem Quadratmittel-Klassifikator überlegen ist. Dies unterstützt die Behauptung aus Abschnitt 4.2, daß die KL-Koeffizienten annähernd normal verteilt sind.

Anhang C und D enthalten die Klassifikationsergebnisse nach Anwendung der ermittelten optimalen, linearen Transforma-

tionen. Anhang C gruppiert die Ergebnisse nach Zieldimension n', Anhang D faßt die Ergebnisse eines festen Klassifikators zusammen.

Die wesentlichen Resultate können aus den Bildern 4-2 und 4-3 abgelesen werden.

Die Bilder 4-2 und 4-3 zeigen die Klassifikationsleistung des Normalverteilungsklassifikators in Abhängigkeit von der Anzahl der verwendeten Merkmale für die beiden verwendeten Datensätze. Die Tabellen in Anhang D zeigen, daß die entsprechenden Diagramme für die anderen beiden Klassifikatoren prinzipiell den gleichen Aufbau zeigen. Es ergibt sich lediglich eine Verschiebung der Kurven nach oben, d. h., die beiden anderen Klassifikatoren arbeiten insgesamt etwas schlechter.

Grundsätzlich geht aus den Bildern 4-2 und 4-3 hervor, daß die Merkmale, die mit Hilfe der klassifikatorbezogenen Merkmalsauswahl gewonnen wurden, die besten Klassifikationsergebnisse liefern (Kurve MMA-optimale Merkmale). Die Optimierung der Divergenz verbessert die Klassifikationsleistung nicht gegenüber der Verwendung der gleichen Anzahl von KL-Koeffizienten, in einigen Fällen führt die Optimierung der Divergenz sogar auf erhebliche Verschlechterungen in der Fehlerrate. Dies ist zum Beispiel für die Schriftzeichen bei der Reduktion von 30 auf 15 Merkmale der Fall, wo die Fehlerrate des Normalverteilungsklassifikators von 1,91 % für die KL-Koeffizienten auf 2,39 % ansteigt.

Es fällt auf, daß die iterierte Hauptachsentransformation Merkmale liefert, die eine Klassifikationsleistung liefern, die gegenüber den anderen Merkmalsauswahlverfahren deutlich abfällt. Die Einbeziehung Trennflächen höherer als zweiter Ordnung bringt bei den verwendeten Klassifikatoren also nicht die erhoffte Verbesserung in der Fehlerrate. Die beobachtete Verschlechterung hingegen zeigt deutlich, daß die

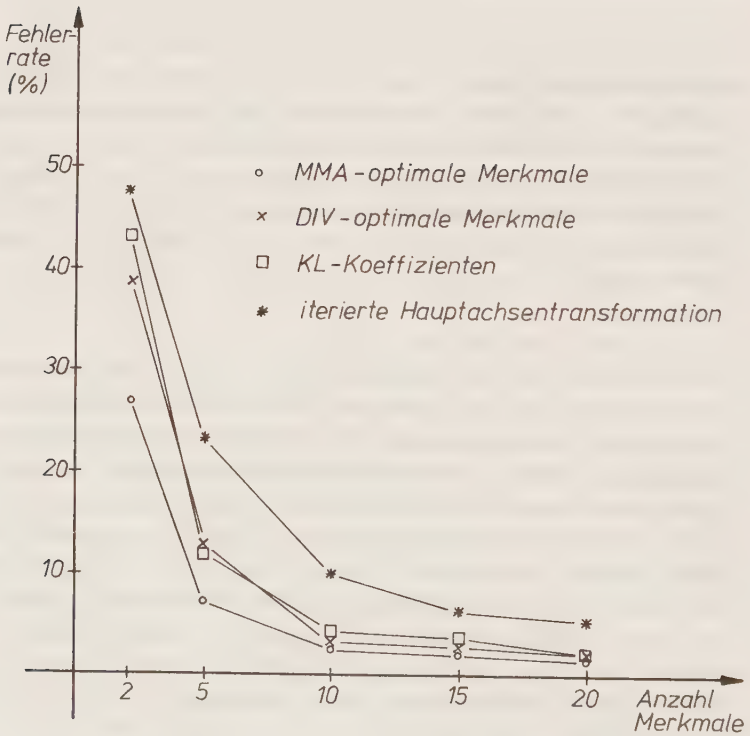


Bild 4-2 Leistung des Normalverteilungsklassifikators als Funktion der Merkmalszahl am Beispiel Sprache

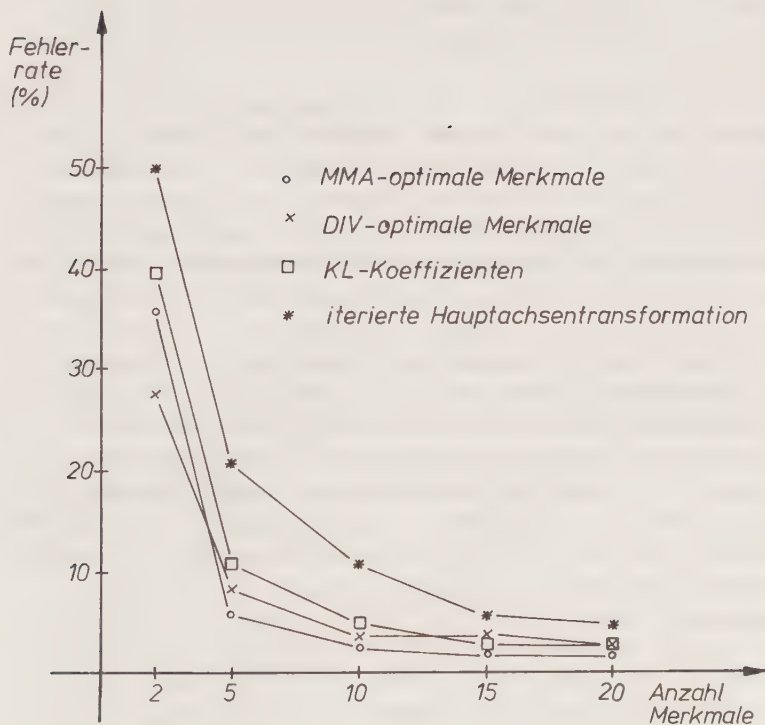


Bild 4-3 Leistung des Normalverteilungsklassifikators als Funktion der Merkmalszahl am Beispiel Schrift

KL-Transformation nicht in der Lage ist, gute Merkmale zu erzeugen. Die KL-Transformation kann also nur als Dimensionsreduktionsverfahren dienen, wenn geometrische Rekonstruierbarkeit erreicht werden soll. Die Beibehaltung von Klassifikationsleistungen gelingt nur bei fast perfekter geometrischer Rekonstruierbarkeit. Für die vorliegenden Versuche wurde das vollständige quadratische Polynom aus 20 KL-Koeffizienten zugrunde gelegt.

Das gute Abschneiden der MMA-optimalen Merkmale soll anhand einiger Zahlen belegt werden. Zunächst zeigt Anhang D, daß für Merkmalszahlen 5 und größer die MMA-optimalen Merkmale für alle drei Klassifikatoren ausnahmslos die besten Klassifikationsergebnisse liefern. Dies demonstriert bereits die Brauchbarkeit des Konzepts "klassifikatorbezogene Merkmalsauswahl". Die erreichbare Verbesserung läßt sich zum einen, wie in Anhang C bzw. D getan, durch Vergleich von Fehlerraten darstellen. Zum anderen ist jedoch auch interessant, wieviele Merkmale sich bei gleicher Fehlerrate einsparen lassen, wenn man MMA-optimale Merkmale verwendet. Die Tabellen 4-5 und 4-6 zeigen die Anzahl KL-Koeffizienten, die die gleiche Klassifikationsleistung beim Normalverteilungsklassifikator erbringen wie die dazu angegebene Zahl MMA-optimaler Merkmale. Die Anzahlen von KL-Koeffizienten wurden durch lineare Interpolation in den Tabellen von Anhang B gewonnen.

Vergleich von Merkmalszahlen am Beispiel Sprache

Anzahl MMA-optimale Merkmale	Fehlerrate des NVK	äquivalente Anzahl KL-Koeffizienten
5	7,35 %	7
10	2,65 %	16
15	2,37 %	18
20	1,49 %	26

Tabelle 4-5

Vergleich von Merkmalszahlen am Beispiel Schrift

Anzahl MMA-optimale Merkmale	Fehlerrate des NVK	äquivalente Anzahl KL-Koeffizienten
5	6,30 %	7
10	2,13 %	15
15	1,73 %	18
20	1,51 %	23

Tabelle 4-6

Der Vergleich zeigt, daß man bei geeigneter Merkmalsauswahl bei gleicher Klassifikationsleistung statt mit 26 Koeffizienten auch mit 20 Merkmalen auskommen kann.

Für Reduktionen mit Zieldimension 2 können die räumlichen Gegebenheiten mit Hilfe der klassenäquivalenten Ellipsen /S1/ veranschaulicht werden. Anhang E zeigt die klassenäquivalenten Ellipsen für die jeweils zwei besten Merkmale nach jeder der fünf Auswahlmethoden. Die gezeigten Ellipsen genügen jeweils Gleichung (3-13), wenn die Konstante "const" den Wert 2 besitzt.

Eine letzte Frage ist, inwieweit die Anzahl n der Eingangsmerkmale das erreichbare Klassifikationsergebnis beeinflußt. Ein entsprechender Vergleich wurde für Zieldimension $n' = 2$ angestrebt. Es wurde Eingangsdimension 30 mit Eingangsdimension 70 verglichen. Anhang A zeigt, daß die Werte der Gütemaße in allen Fällen besser werden. Diese Verbesserung des Wertes des Gütemaßes führt jedoch nicht immer zu einer Verbesserung in der Klassifikationsleistung. So zeigen die NVK-optimale Merkmale einen deutlichen Überanpassungseffekt: während die Fehlerrate bei den Sprachdaten für die Lernstichprobe (= Wert des Gütemaßes) beim Übergang von 30 auf 70 Eingangsmerkmale von 16,27 % auf 15,64 % gesenkt wird, steigt

die an der Teststichprobe gemessene Fehlerrate des Normalverteilungsklassifikators von 21,62 % auf 23,10 % an. Das gleiche Phänomen zeigt die Divergenz für die Schriftzeichen, wobei hier die Fehlerrate des Normalverteilungsklassifikators von 27,86 % auf 32,50 % ansteigt. Eine uneingeschränkte Verbesserung bringen nur die MMA-optimalen Merkmale. Diese Merkmale verbessern die Klassifikationsleistung des Normalverteilungsklassifikators am Beispiel Sprache von 26,80 % auf 23,14 % und am Beispiel Schrift von 36,37 % auf 34,74 %.

5. Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit dem Problem der linearen Merkmalsauswahl für normalverteilte Merkmale. Nach einer Übersicht über bekannte Verfahren wurde das Konzept der "klassifikatorbezogenen Merkmalsauswahl" vorgestellt. Der wesentliche Gedanke dieses Konzeptes ist, eine Brücke zwischen den klassischen Teilen eines Klassifikationssystems, nämlich der Merkmalsextraktionsstufe und dem eigentlichen Klassifikator, zu schlagen. Das Konzept fordert, daß der Entwurf der Merkmalsextraktionsstufe sich an der Struktur des zu verwendeten Klassifikators zu orientieren hat. Die klassische Lösung sieht den Entwurf einer Merkmalsextraktionsstufe gänzlich unabhängig von dem verwendeten Klassifikator.

Nach der Einführung des Konzepts wird an zwei Beispielen untersucht, wie man zu Merkmalen kommt, die sich für einen vorgegebenen Klassifikator gut eignen. Als Beispiele werden der Normalverteilungsklassifikator und der modifizierte Minimum-Abstand-Klassifikator gewählt (NVK bzw. MMA genannt). Zur Bestimmung der besten linearen Merkmale für den NVK wird die Bayesfehlerrate als Funktion der gesuchten Transformationsmatrix ausgedrückt. Die Auswertung der resultierenden Funktion ist wegen der mehrfachen Integrationen mit nur impliziert festgelegten Integrationsgrenzen weder analytisch noch numerisch durchführbar. Im Experiment muß deshalb die Fehlerrate mit Hilfe einer Lernstichprobe geschätzt werden. Dieser Vorgang erweist sich als äußerst rechenaufwendig.

Das Konzept wird weiterhin auf den MMA angewandt. Das zu optimierende Gütekriterium ist ein Schätzwert für die Fehlerwahrscheinlichkeit. Geometrisch bedeutet das Kriterium das Aufsuchen von Hyperellipsoiden, die den vom MMA induzierten Klassenbereichen einbeschrieben sind. Das Verfahren sucht eine lineare Merkmalsauswahlfunktion, welche diese Hyperellipsoide möglichst groß macht.

Im experimentellen Teil wird ein Vergleich zwischen fünf verschiedenen Merkmalsauswahlverfahren durchgeführt. Es sind dies die

- i) lineare Merkmalsauswahl nach der Karhunen-Loève-Transformation (KL-Koeffizienten)
- ii) lineare Merkmalsauswahl durch Maximierung der mittleren Divergenz (DIV-optimale Merkmale)
- iii) lineare Merkmalsauswahl durch Minimierung der Fehler-rate beim MMA (MMA-optimale Merkmale)
- iv) lineare Merkmalsauswahl durch Minimierung der Fehler-rate beim NVK (NVK-optimale Merkmale)
- v) Merkmalsauswahl durch iterierte Hauptachsentransformation

Die Untersuchungen an zwei verschiedenen Stichproben (Sprachdaten und Schriftzeichen, jeweils die zehn Klassen "0" bis "9") ergaben ein einheitliches Bild von der Leistungsfähigkeit der fünf Auswahlverfahren. Demnach sind die MMA-optimale Merkmale, welche mit Hilfe des Konzeptes der klassifikatorbezogenen Merkmalsauswahl gewonnen wurden, allen anderen Auswahlverfahren überlegen. Die enorme Leistungsfähigkeit des Auswahlverfahrens zeigt sich zum Beispiel bei der Reduktion von 30 auf 5 Merkmale, wo die MMA-optimale Merkmale für die Sprachdaten bei Verwendung des NVK gegenüber den KL-Koeffizienten eine Verbesserung von 12,29 % auf 7,35 % bringen. Bei den Schriftzeichen wird eine Verbesserung von 11,6 % auf 6,3 % erreicht.

Die Verwendung guter Merkmale kann auch zur Verringerung der Anzahl der zur Einhaltung einer vorgegebenen Leistung notwendigen Anzahl von Merkmalen verwendet werden. Die MMA-optimale Merkmale erweisen sich auch in dieser Hinsicht als sehr leistungsfähig. So kann bei den Sprachdaten die Leistungsfähigkeit von 26 KL-Koeffizienten bereits durch

20 MMA-optimale Merkmale erbracht werden.

Nachdem durch die Experimente zur vorliegenden Arbeit das Konzept der klassifikatorbezogenen Merkmalsauswahl als tragfähig erwiesen wurde, stellt sich die Frage, ob das Konzept für beliebige Klassifikatoren erweitert werden kann. Eine Entscheidung hierüber muß sicherlich weiteren Untersuchungen vorbehalten bleiben. Insbesondere wird noch einige Entwicklungsarbeit zu leisten sein, um die zur Bestimmung klassifikatorbezogener Merkmale erforderlichen Rechnungen durchführen zu können. In jedem Fall kann man hoffen, daß auf dem Weg der klassifikatorbezogenen Merkmalsauswahl weitere Fortschritte zur Lösung des Problems Merkmalsauswahl in der Mustererkennung erreicht werden.

Literaturverzeichnis

Abkürzungen:

IEEE Institute of Electrical and Electronics Engineers
IJCPR International Joint Conference on Pattern Recognition

- /A1/ Andrews, H.:
Multidimensional Rotations in Feature Selection
IEEE Transactions on Computers, Vol. C-20, 1971,
S. 1046-1056
- /B1/ Becker, D.:
Vergleich eines lineären und eines nichtlinearen
Klassifikators bei der Worterkennung
Wissenschaftliche Berichte AEG-Telefunken 47, 1974,
S. 77-84
- /B2/ Björk, Å.; Dahlquist, G.:
Numerische Methoden
Oldenbourg Verlag, München-Wien, 1972
- /C1/ Cover, T.M.; Hart, P.E.:
Nearest Neighbour Pattern Classification
IEEE Transactions on Information Theory, Vol. IT-13,
1967, S. 21-27
- /C2/ Chen, C.H.:
On Information and Distance Measures, Error Bounds,
and Feature Selection
Information Sciences, Vol. 10, 1976, S. 159-173
- /C3/ Chang, C.:
Dynamic Programming as applied to Feature Subset
Selection in a Pattern Recognition System
IEEE Transactions on Systems, Man and Cybernetics,
Vol. SMC-3, 1973, S. 166-171

- /C4/ Cover, T.M.:
The Best Two Independent Measurements are not the
Best Two
IEEE Transactions on System Science and Cybernetics,
Vol. SSC-10, 1974, S. 116-117
- /D1/ Duda, R.O.; Hart, P.E.:
Pattern Classification and Scene Analysis
John Wiley and Sons, Inc., New York, 1973
- /D2/ Decell, H.P.; Quirein, J.A.:
An iterative Approach to the Feature Selection
Problem
Proc. Machine Processing of Remotely Sensed Data,
West Lafayette, USA, Oktober 1973, S. 3B1-3B12
- /E1/ Elashoff, J.D.:
On the choise of variables in classification prob-
lems with dichotomous variables
Biometrika, Vol. 54, 1967, S. 668-670
- /F1/ Fukunaga, K.:
Introduction to Statistical Pattern Recognition
Academic Press, New York, 1972
- /F2/ Fu, K.S.:
Sequential Methods in Pattern Recognition and Ma-
chine Learning
Academic Press, New York, 1968
- /F3/ Fu, K.S.:
Syntactic Methods in Pattern Recognition
Academic Press, New York, 1974

- /F4/ Fukunaga, K.; Koontz, L.G.:
Representation of Random Processes using finite
Karhunen-Loève-expansion
Information and Control, Vol. 16, 1970, S. 85-101
- /F5/ de Figueiredo, R.J.P.:
Optimal Linear and Nonlinear Feature Extraction
Based on the Minimization of the Increased Risk of
Misclassification
Rice University, ICSA-Report Nr. 275-025-014,
Juni 1974
- /F6/ de Figueiredo, R.J.P. et al.:
An algorithm for extraction of more than one opti-
mal linear Feature from several Gaussian Pattern
Classes
Proc. 4. IJCPR, Kyoto, Japan, 1978, S. 793-797
- /F7/ Foley, D.H.:
Considerations of Sample and Feature Size
IEEE Transactions on Information Theory, Vol. IT-
18, 1972, S. 618-622
- /F8/ Fletcher R.; Powell, J.:
A Rapidly Converging Descent Method of Minimization
British Computer Journal, 1963, S. 163-168
- /H1/ Ho, Y.C.; Agrawala, A.K.:
On pattern classification algorithms - introduction
and survey
IEEE Transactions on Automatic Control, Vol. AC-13,
1968, S. 676-690

- /H2/ Hughes, G.F.:
On the Mean Accuracy of Statistical Pattern Recognizers
IEEE Transactions on Information Theory, Vol. IT-14, 1968, S. 55-63
- /H3/ Horst, R.:
Nichtlineare Optimierung
Hanser Verlag, München-Wien, 1979
- /H4/ Highleyman, W.H.:
The design and analysis of pattern recognition experiments
The Bell System Technical Journal, 1962, S. 723-744
- /J1/ Jäpel, D.:
Merkmalsauswahl in der Mustererkennung
Arbeitsbericht am Lehrstuhl für Informatik 5
(Mustererkennung) der Universität Erlangen, 1977
- /K1/ Kanal, L.:
Patterns in Pattern Recognition: 1968-1974
IEEE Transactions on Information Theory, Vol.IT-20, 1974, S. 697-722
- /K2/ Kittler, J.:
Mathematical Methods of Feature Selection in Pattern Recognition
International Journal of Man-Machine Studies, Vol. 7, 1975, S. 609-637
- /K3/ Kulkarni, A.V.:
An Optimization Approach to Hierarchical Classifier design
Proc. 3. IJCPR, Coronado, USA, 1976, S. 459-466

- /L1/ Luenberger, D.G.:
Optimization by Vector Space Methods
John Wiley and Sons, Inc., New York, 1969
- /M1/ Mucciardi, A.N.; Gose, E.E.:
A Comparison of Seven Techniques for Choosing Sub-
sets of Pattern Recognition Properties
IEEE Transactions on Computers, Vol. C-20, 1971
S. 1023-1031
- /M2/ von Mangoldt; Knopp.:
Einführung in die höhere Mathematik,
Band IV
Hirzel Verlag, Stuttgart, 1973
- /N1/ Niemann, H.:
Mustererkennung - Anwendungen
Informatik Spektrum, Band 3, Heft 1, Januar 1980,
S. 19-30
- /N2/ Niemann, H.:
Methoden der Mustererkennung
Akademische Verlagsgesellschaft, Frankfurt, 1974
- /N3/ Niemann, H.:
Mustererkennung mit orthonormalen Reihenentwick-
lungen
Nachrichtentechnische Zeitschrift, Band 23, 1970,
S. 308-313
- /N4/ Niemann, H.; Weiß, J.:
A fast-converging algorithm for nonlinear mapping
of highdimensional data to a plane
IEEE Transactions on Computers, Vol. C-28, 1979,
S. 142-147

- /O1/ Otterloo, P.J. van; Young, I.T.:
A distribution-free geometric upper bound for the
probability of error of a minimum distance classi-
fier
Pattern Recognition, Vol. 10, 1978, S. 281-286
- /O2/ Ott, R.:
Über zweistufige quadratische Klassifikatoren
Dissertation, Universität Erlangen, 1977
- /P1/ Patrick, E.A.:
Fundamentals of Pattern Recognition
Prentice-Hall, Englewood Cliffs, New Jersey, 1972
- /R1/ Reinhardt, K.-H.:
Algorithmen zur Merkmalsauswahl mit Implementierung
eines Verfahrens
Diplomarbeit im Fach Informatik, Universität Er-
langen, 1979
- /S1/ Schürmann, J.:
Polynomklassifikatoren für die Zeichenerkennung
Oldenbourg Verlag, München, 1977
- /T1/ Toussaint, G.T.:
Bibliography on Estimation of misclassification
IEEE Transactions on Information Theory, Vol.
IT-20, 1974, S. 472-479
- /T2/ Tou, J.T.; Gonzales, R.C.:
Pattern Recognition Principles
Addison-Wesley, London, 1974
- /V1/ Vilmansen, T.R.:
Feature Evaluation with Measures of Probabilistic
Dependence
IEEE Transactions on Computers, Vol. C-22, 1973,
S. 381-387

- /W1/ Watanabe, S.:
Karhunen-Loève Expansion and Factor Analysis
Transactions 4.th Prague Conference on Information
Theory, 1965, S. 635-660
- /W2/ Watanabe, S. et al.:
Evaluation and Selection of Variables in Pattern
Recognition
in: Tou, J.T. (Hrsg.): Computer and Information
Sciences II
Academic Press, New York, 1967, S. 92-122
- /W3/ Wilks, S.:
Mathematical Statistics
John Wiley and Sons, Inc., New York, 1962
- /Y1/ Young, T.Y.; Calvert, T.W.:
Classification, Estimation and Pattern Recognition
Elsevier, New York, 1974
- /Z1/ Zurmühl, R.:
Matrizen und ihre technischen Anwendungen
Springer, Berlin, 1964 (4. Auflage)

Anhang A

Werte der Gütemaße bei der Berechnung der Transformationsmatrizen für

- . MMA-optimale
- . DIV-optimale
- . NVK-optimale

Merkmale.

In den Tabellen ist der Wert der Gütemaße jeweils für n KL-Koeffizienten (unter Abschnitt A) und für die n Merkmale angegeben, die das zugehörige Gütemaß optimieren (unter Abschnitt B). Da als Startmatrix in allen Optimierungsläufen die Einheitsmatrix verwendet wurde, gibt der für m KL-Koeffizienten angegebene Wert den Startwert des Gütemaßes bei einer Reduktion auf m Merkmale an.

- A) Wert des MMA-Kriteriums für die ersten KL-Koeffizienten bei den Sprachdaten

Anzahl	Wert
2	10,729
5	2,751
10	1,921
15	2,088
20	1,981
30	2,261

- B) Wert des MMA-Kriteriums für MMA-optimale Merkmale bei den Sprachdaten

1) Ausgangsdimension 30

Anzahl	Wert
2	1,407
5	1,160
10	1,342
15	1,619
20	1,654

2) Ausgangsdimension 70

Anzahl	Wert
2	1,110

- A) Wert des NVK-Kriteriums (geschätzte Fehlerwahrscheinlichkeit, Schätzung mit Hilfe der Lernstichprobe) für die ersten KL-Koeffizienten bei den Sprachdaten

Anzahl	Wert
2	42,02
5	11,16

- B) Wert des NVK-Kriteriums für NVK-optimale Merkmale bei den Sprachdaten

1) Ausgangsdimension 30

Anzahl	Wert
2	16,27
5	7,403

2) Ausgangsdimension 70

Anzahl	Wert
2	15,64

- A) mittlere Divergenz über alle Klassenpaare für die ersten KL-Koeffizienten bei den Sprachdaten

Anzahl	Wert
2	24,94
5	85,04
10	173,35
15	268,26
20	367,12
30	715,75
70	3047,06

- B) mittlere Divergenz über alle Klassenpaare für DIV-optimale Merkmale bei den Sprachdaten

1) Ausgangsdimension 30

Anzahl	Wert
2	167,26
5	211,75
10	378,38
15	467,52
20	558,86

2) Ausgangsdimension 70

Anzahl	Wert
2	227,27

- A) Wert des MMA-Kriteriums für die ersten KL-Koeffizienten bei der Stichprobe Schrift

Anzahl	Wert
2	106,27
5	4,73
10	1,84
15	1,85
20	2,02
30	2,54

- B) Wert des MMA-Kriteriums für MMA-optimale Merkmale bei der Stichprobe Schrift

- 1) Ausgangsdimension 30

Anzahl	Wert
2	2,33
5	1,01
10	1,09
15	1,44
20	1,82

- 2) Ausgangsdimension 70

Anzahl	Wert
2	2,10

- A) Wert des NVK-Kriteriums (geschätzte Fehlerwahrscheinlichkeit, Schätzung mit Hilfe der Lernstichprobe) für die ersten KL-Koeffizienten bei der Stichprobe Schrift.

Anzahl	Wert
2	41,77
5	12,85

- B) Wert des NVK-Kriteriums für NVK-optimale Merkmale bei der Stichprobe Schrift

1) Ausgangsdimension 30

Anzahl	Wert
2	32,83
5	6,82

2) Ausgangsdimension 70

Anzahl	Wert
2	32,83

- A) mittlere Divergenz über alle Klassenpaare für die ersten KL-Koeffizienten bei der Stichprobe Schrift

Anzahl	Wert
2	22,23
5	46,12
10	94,46
15	144,46
20	189,50
30	271,74
70	578,26

- B) mittlere Divergenz über alle Klassenpaare für DIV-optimale Merkmale

1) Ausgangsdimension 30

Anzahl	Wert
2	49,44
5	90,59
10	150,06
15	190,95
20	223,60

2) Ausgangsdimension 70

Anzahl	Wert
2	55,97

Anhang B)

Klassifikationsergebnisse bei Verwendung der KL-Koeffizienten als Merkmale.

Angegeben ist die Leistung des Normalverteilungsklassifikators (NVK-Leistung), des modifizierten Minimum-Abstand-Klassifikators (MMA-Leistung) und des Quadratmittel-Klassifikators (QMK-Leistung), jeweils der prozentuale Anteil falsch klassifizierter Muster der Teststichproben. Die QMK-Leistung wird mit Hilfe des vollständigen quadratischen Polynoms der Eingangsmerkmale ermittelt.

Für mehr als 20 Merkmale wurde die QMK-Leistung nicht ermittelt. Der Grund hierfür ist, daß die Kovarianzmatrix des vollständigen quadratischen Polynoms die Speicherkapazität der vorhandenen Rechenanlage übersteigt.

Stichprobe Sprache

Leistungen in %-Fehlerrate

Anzahl KL-Koeffizienten	NVK- Leistung	MMA- Leistung	QMK- Leistung
2	43,15	42,07	50,88
3	27,28	27,10	37,57
4	21,74	22,68	27,40
5	12,29	13,17	17,56
6	9,14	10,42	12,57
7	7,47	7,97	8,53
8	6,29	6,63	7,09
9	5,16	5,66	5,99
10	4,06	4,02	4,46
15	2,93	3,38	3,34
20	1,93	2,45	2,15
25	1,53	2,09	x
30	1,35	1,73	x
35	1,17	1,73	x
40	1,21	1,61	x
45	0,99	1,41	x
50	1,11	1,39	x
55	1,11	1,39	x
60	1,15	1,41	x
65	1,17	1,43	x
70	1,15	1,49	x

Stichprobe Schrift

Leistungen in %-Fehlerrate

Anzahl KL-Koeffizienten	NVK- Leistung	MMA- Leistung	QMK- Leistung
2	39,93	41,41	49,10
3	26,79	27,35	34,21
4	19,60	20,46	25,05
5	11,66	12,33	16,29
6	9,14	9,83	12,61
7	4,56	5,02	6,72
8	4,63	4,96	6,28
9	4,14	4,62	5,49
10	3,00	3,28	4,07
15	1,91	2,00	2,25
20	1,62	1,73	1,55
25	1,40	1,53	x
30	1,25	1,42	x
35	1,20	1,33	x
40	1,23	1,35	x
45	1,25	1,42	x
50	1,36	1,51	x
55	1,40	1,56	x
60	1,46	1,60	x
65	1,49	1,66	x
70	1,48	1,70	x

Anhang C

Ergebnisse der Klassifikation nach Anwendung der berechneten linearen Transformationen

Auf jeweils einer Seite sind die Ergebnisse zu einer festen Reduktion von n auf n' Merkmale zusammengefaßt. Die Leistung der Klassifikatoren sind in %-Fehlerrate auf der Teststichprobe angegeben. Die Abkürzungen in den Kopfzeilen der Tabellen bedeuten:

MMA	MMA-optimale Merkmale
NVK	NVK-optimale Merkmale
DIV	DIV-optimale Merkmale
KL	KL-Koeffizienten
ITERKL	iterierte Hauptachsentransformation, vollständig quadratisches Polynom in 20 KL-Koeffizienten

Die Zahlen zu einigen Plätzen in den Tabellen wurden nicht ermittelt. Die Gründe hierfür sind:

1. Die QMK-Leistung für die iterierte Hauptachsentransformation wurde wegen des schlechten Abschneidens dieser Merkmale grundsätzlich nicht ermittelt.
2. Die NVK-optimalen Merkmale für Reduktionen mit Zieldimension größer als 5 wurden wegen des hohen Rechenzeitbedarfs nicht ermittelt. (Siehe auch Abschnitt 4.4.3.3).
3. Wegen technischer Defekte mußte auf die Auswertung der QMK-Leistung für die Sprachdaten bei Zieldimension 2, 15 und 20 verzichtet werden.

Reduktion 30 auf 2 Merkmale

A) NVK-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	26,80	21,62	38,17	43,15	47,35
Schrift	36,37	36,12	27,86	39,93	50,56

B) MMA-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	27,16	22,82	41,77	42,07	48,85
Schrift	36,84	36,48	28,90	41,41	52,44

C) QMK-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	x	x	x	50,88	x
Schrift	50,36	46,23	50,35	49,10	x

Reduktion 30 auf 5 Merkmale

A) NVK-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	7,35	11,50	12,46	12,29	23,06
Schrift	6,30	8,73	7,63	11,60	21,06

B) MMA-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	7,77	12,60	15,65	13,17	24,79
Schrift	6,66	8,96	8,54	12,30	20,35

C) QMK-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	11,85	16,76	16,64	17,56	x
Schrift	8,69	12,66	12,58	16,29	x

Reduktion 30 auf 10 Merkmale

A) NVK-Leistung

	MMA	NVK	DEV	KL	ITERAL
Sprache	2,65	x	3,55	4,05	5,34
Schrift	2,73	x	2,95	3,00	4,35

B) MMA-Leistung

	MMA	NVK	DEV	KL	ITERAL
Sprache	3,17	x	5,01	4,01	5,52
Schrift	2,19	x	3,26	3,28	12,25

C) QMK-Leistung

	MMA	NVK	DEV	KL	ITERAL
Sprache	3,95	x	5,17	4,45	x
Schrift	2,85	x	3,81	4,07	x

Reduktion 30 auf 15 Merkmale

A) NVK-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	2,37	x	2,67	2,93	6,43
Schrift	1,73	x	2,39	1,91	5,72

B) MMA-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	2,83	x	3,53	3,38	6,45
Schrift	1,82	x	2,50	2,00	6,52

C) QMK-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	x	x	x	3,34	x
Schrift	2,18	x	2,66	2,25	x

Reduktion 30 auf 20 Merkmale

A) NVK-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	1,49	x	1,97	1,93	5,42
Schrift	1,51	x	1,79	1,62	4,15

B) MMA-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	1,87	x	2,49	2,45	5,22
Schrift	1,60	x	1,98	1,73	4,28

C) QMK-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	x	x	x	2,15	x
Schrift	1,62	x	1,81	1,55	x

Reduktion 70 auf 2 Merkmale

A) NVK-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	23,14	23,10	26,39	43,15	47,35
Schrift	34,74	36,12	32,50	39,93	50,56

B) MMA-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	23,16	23,56	27,37	42,07	48,85
Schrift	35,31	36,49	33,08	41,41	52,44

C) QMK-Leistung

	MMA	NVK	DIV	KL	ITERKL
Sprache	x	x	x	50,88	x
Schrift	48,68	46,23	50,50	49,10	x

Anhang D

Ergebnisse der Klassifikation nach Anwendung der berechneten linearen Transformation

Jede der hier angeführten Tabellen faßt die Klassifikationsergebnisse für einen der drei Klassifikatoren zusammen. Die Spalten der Tabellen sind mit den Namen der Merkmalsauswahlverfahren beschriftet. Es werden die in Anhang C angegebenen Abkürzungen verwendet. Die Zeilen der Tabellen sind mit der Anzahl der verwendeten Merkmale beschriftet.

Die Zahlen zu einigen Plätzen in den Tabellen wurden nicht ermittelt. Die Gründe werden auf Seite 119 (Anhang C) angegeben.

Stichprobe Sprache

Klassifikationsergebnis des Normalverteilungsklassifikators
in %-Fehlerrate

	MMA	NVK	DIV	KL	ITERKL
2	26,80	21,62	38,17	43,15	47,35
5	7,35	11,50	12,46	12,29	23,06
10	2,65	x	3,85	4,06	9,94
15	2,37	x	2,67	2,93	6,43
20	1,49	x	1,97	1,93	5,42

Stichprobe Sprache

Klassifikationsergebnis des modifizierten Minimum-Abstand-Klassifikators in %-Fehlerrate

	MMA	NVK	DIV	KL	ITERKL
2	27,16	22,82	41,77	42,07	48,85
5	7,77	12,60	15,65	13,17	24,79
10	3,17	x	5,02	4,02	9,82
15	2,83	x	3,53	3,38	6,45
20	1,87	x	2,49	2,45	5,22

Stichprobe Sprache

Klassifikationsergebnis des Quadratmittel-Klassifikators
in %-Fehlerrate

	MMA	NVK	DIV	KL	ITERKL
2	x	x	x	50,88	x
5	11,85	16,76	16,64	17,56	x
10	3,98	x	6,17	4,46	x
15	x	x	x	3,34	x
20	x	x	x	2,15	x

Stichprobe Schrift

Klassifikationsergebnis des Normalverteilungsklassifikators
in %-Fehlerrate

	MMA	NVK	DIV	KL	ITERKL
2	36,37	36,12	27,86	39,93	50,56
5	6,30	8,73	7,63	11,60	21,06
10	2,13	x	2,98	3,90	11,35
15	1,73	x	2,39	1,91	5,72
20	1,51	x	1,79	1,62	4,15

Stichprobe Schrift

Klassifikationsergebnis des modifizierten Minimum-Abstand-Klassifikators in %-Fehlerrate

	MMA	NVK	DIV	KL	ITERKL
2	36,84	36,48	28,90	41,41	52,44
5	6,66	8,96	8,54	12,30	20,35
10	2,19	x	3,26	3,28	12,25
15	1,82	x	2,50	2,00	6,52
20	1,60	x	1,98	1,73	4,28

Stichprobe Schrift

Klassifikationsergebnis des Quadratmittel-Klassifikators
in %-Fehlerrate

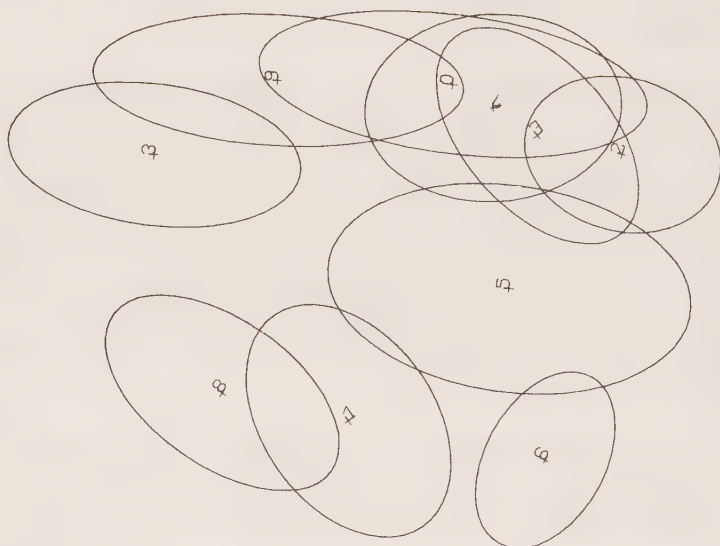
	MMA	NVK	DIV	KL	ITERKL
2	50,36	46,23	50,35	49,10	x
5	8,69	12,66	12,58	16,29	x
10	2,86	x	3,81	4,07	x
15	2,18	x	2,66	2,25	x
20	1,62	x	1,81	1,55	x

Anhang E

Klassenäquivalente Ellipsen für die Reduktionen von 30 auf 2 Merkmale.

Die folgenden Bilder zeigen für die Reduktionen von 30 auf 2 Merkmale die klassenäquivalenten Ellipsen. Diese Ellipsen geben einen Eindruck von der Ausdehnung einer Klasse und vom Grad der Überlappung verschiedener Klassen. Die Ellipsen sind Linien konstanter Wahrscheinlichkeitsdichten. Für Normalverteilungen liefert das Integral über die von den Ellipsen eingeschlossene Fläche in allen Fällen den Wert 0,632; dies bedeutet, daß für klassenweise normalverteilte Merkmale stets ca. 63 % aller Muster einer Klasse innerhalb der gezeigten Ellipsen auftreten.

Auf die Beschriftung der Koordinatenachsen mit Zahlenwerten wurde verzichtet, da die Klassifikationsergebnisse für die hier betrachteten Klassifikatoren unabhängig von Skalierungen sind.

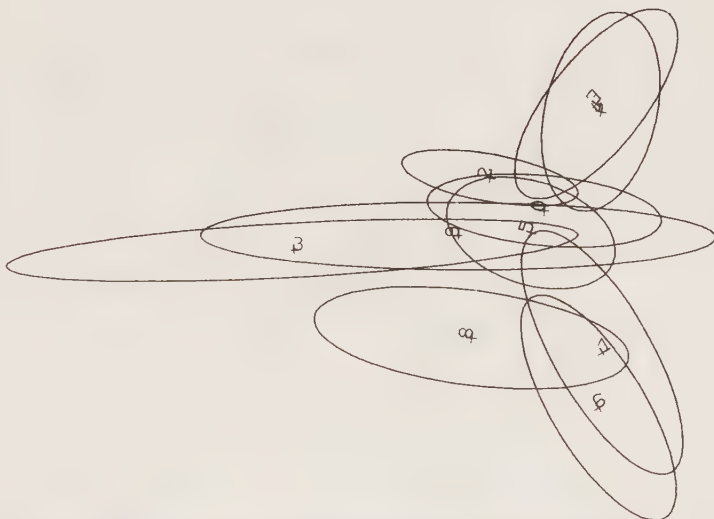


KLASSENÄQUIVALENTE ELLIPSEN

STICHPROBE: SPRACHE

1. UND 2. KL-KOEFFIZIENT

BILD E-1



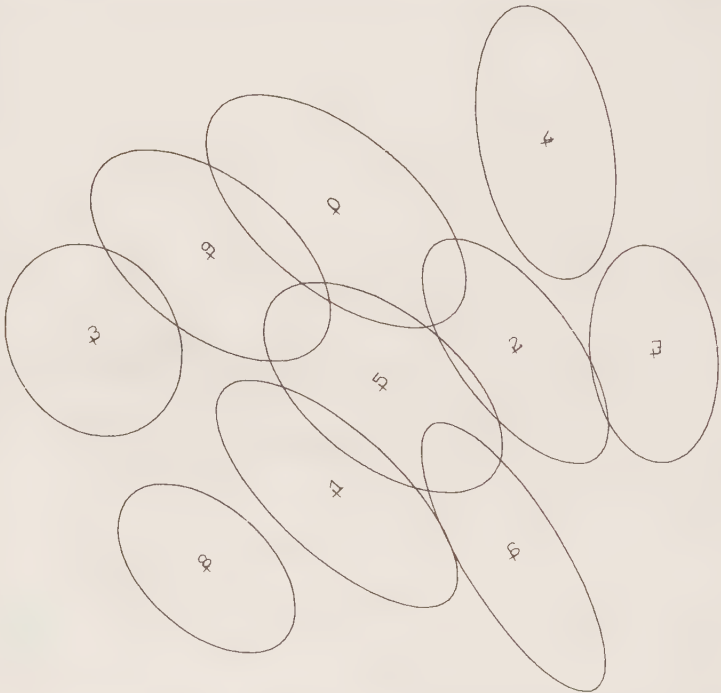
KLASSENÄQUIVALENTE ELLIPSEN

STICHPROBE: SPRACHE

ITERIERTE HAUPTACHSENTTRANSFORMATION

1. UND 2. KOEFFIZIENT

BILD E-2



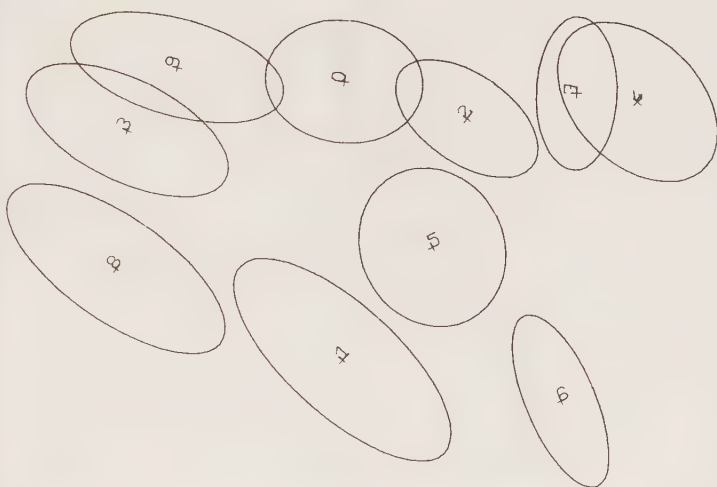
KLASSENÄQUIVALENTE ELLIPSEN

STICHPROBE SPRACHE

MMA - OPTIMALE MERKMALE

REDUKTION 30 AUF 2 MERKMALE

BILD E-3

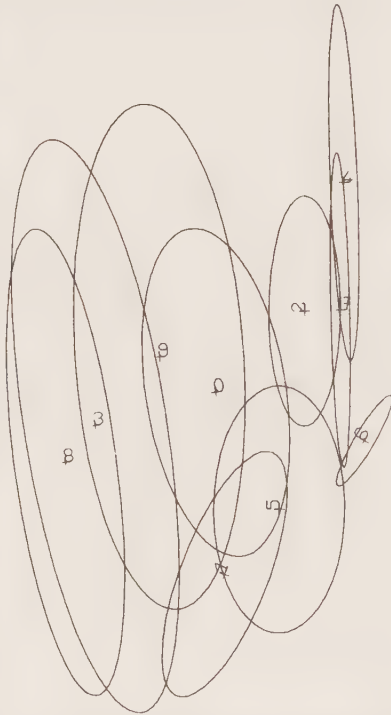


KLASSENÄQUIVALENTE ELLIPSEN

STICHPROBEN SPRACHE

NUK - OPTIMALE MERKMALE

REDUKTION 30 AUF 2 MERKMALE



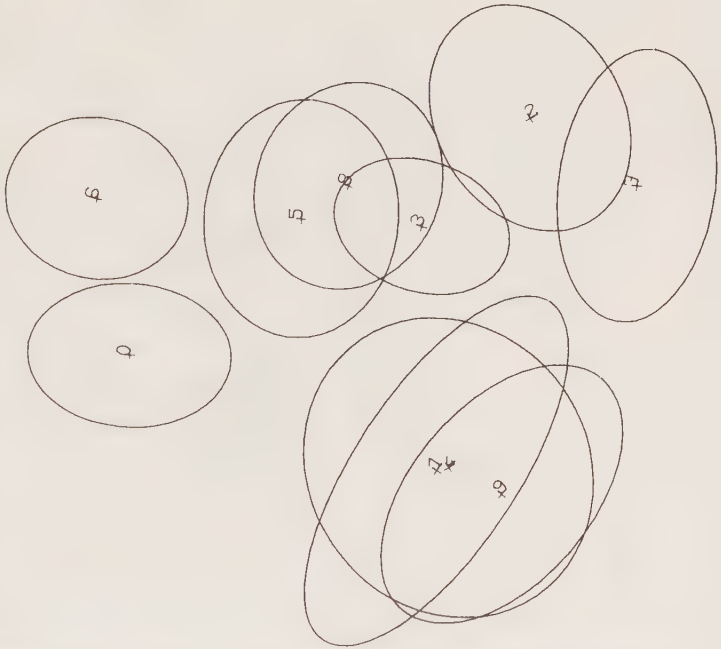
KLASSENÄQUIVALENTE ELLIPSEN

STICHPROBE: SPRACHE

DIV - OPTIMALE MERKMALE

REDUKTION 30 AUF 2 MERKMALE

BILD E-5

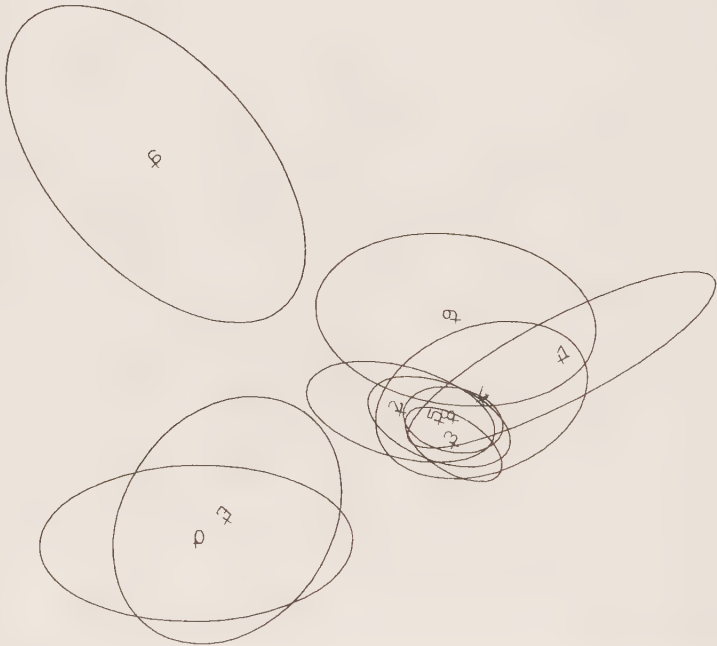


KLASSENÄQUIVALENTE ELLIPSEN

STICHPROBE: SCHRIFT

1. UND 2. KL-KOEFFIZIENT

BILD E-6



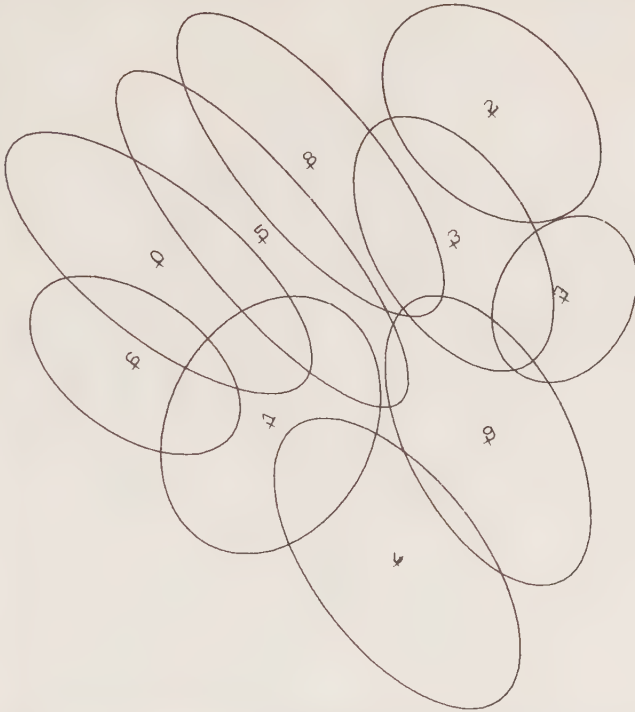
KLASSENÄQUIVALENTE ELLIPSEN

STICHPROBE: SCHRIFT

ITERIERTE HAUPTACHSENTTRANSFORMATION

1. UND 2. KOEFFIZIENT

BILD E-7



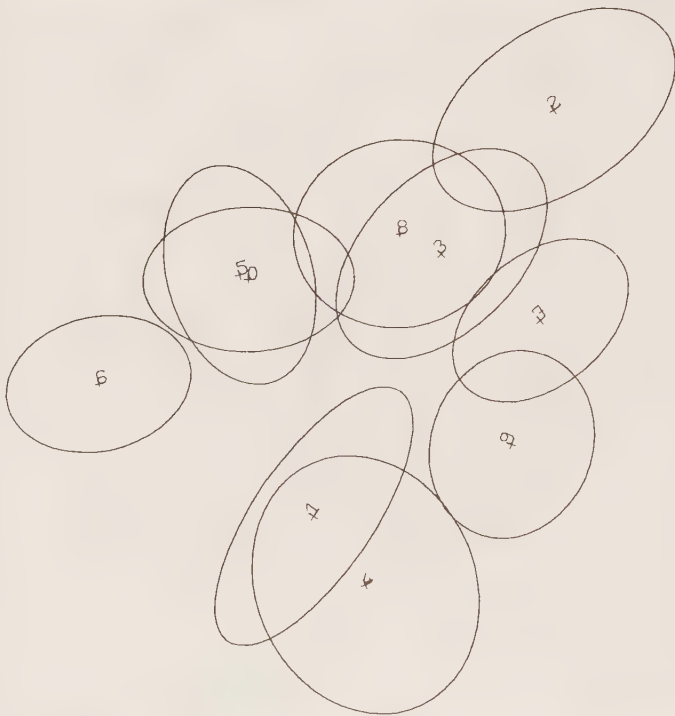
KLASSENÄQUIVALENTE ELLIPSEN

STICHPROBE: SCHRIFT

MMA - OPTIMALE MERKMALE

REDUKTION 30 AUF 2 MERKMALE

BILD E-8

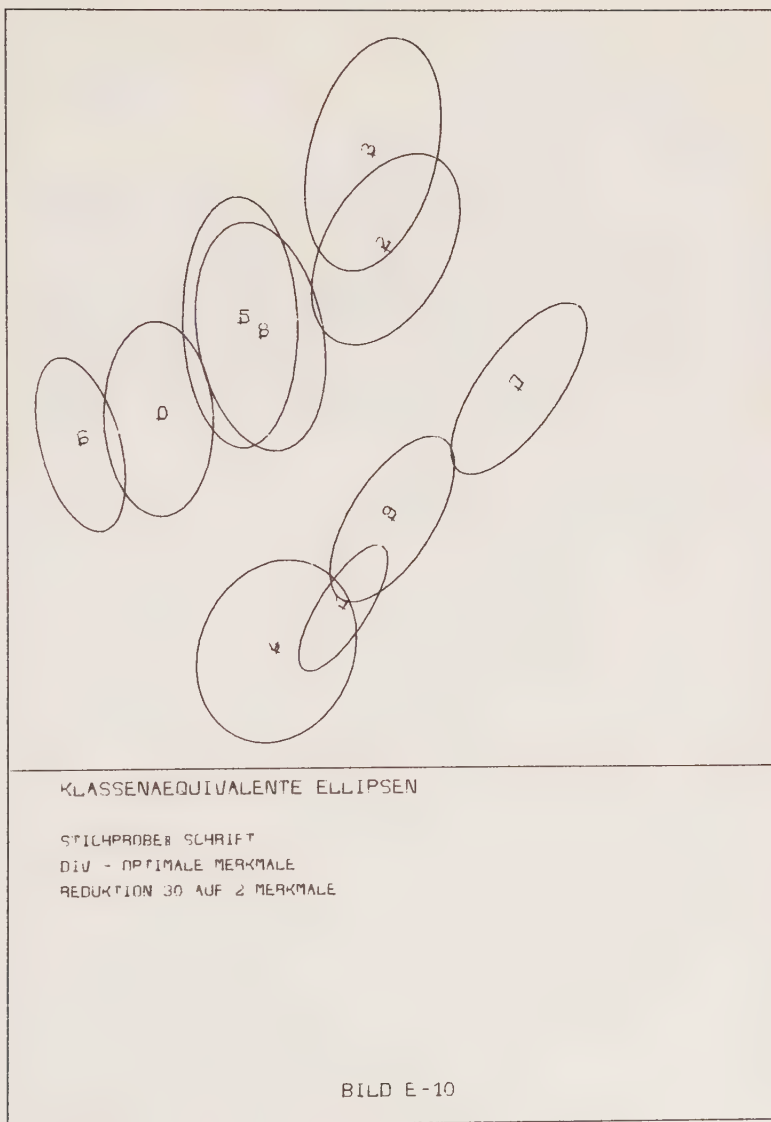


KLASSENÄQUIVALENTE ELLIPSEN

STICHPROBE: SCHRIFT

NUK - OPTIMALE MERKMALE

REDUKTION 30 AUF 2 MERKMALE



Anhang F

Verzeichnis der wichtigsten Symbole

Generell gilt: Vektoren werden mit unterstrichenen kleinen lateinischen Buchstaben, Matrizen mit unterstrichenen großen lateinischen Buchstaben bezeichnet. Ausnahme ist die Bezeichnung für die Mittelwertvektoren. Im Rahmen dieser Arbeit werden Vektoren immer als Spaltenvektoren aufgefaßt.

κ, λ	Klassenindizes
Ω_κ	Klasse κ
$\underline{c}$	Merkmalsvektor
$u_\kappa(\underline{c})$	Prüfgröße zur Klasse Ω_κ
$p(\Omega_\kappa)$	a priori Wahrscheinlichkeit der Klasse Ω_κ
$w(\underline{c}, \Omega_\kappa)$	klassenbedingte Dichte des Merkmalvektors
$\underline{\mu}_\kappa$	Mittelwertvektor des Merkmalvektors für Muster aus Klasse Ω_κ
$\underline{K}_\kappa$	Kovarianzmatrix des Merkmalvektors für Muster aus Klasse Ω_κ
k	Klassenanzahl in einem festen Problem
n	Dimension des Merkmalvektors, insbesondere Ausgangsdimension einer Merkmalsextraktion
n'	Zieldimension einer Merkmalsextraktion
V	Bayesrisiko
$\mathbb{R}^n$	n-dimensional, euklidischer Raum
Ω_κ^O	Bereich des Merkmalraums, der durch den Bayesklassifikator der Klasse Ω_κ zugeordnet wird
$E\{\dots\}$	Erwartungswert einer Zufallsvariablen
$N(\underline{\mu}, \underline{K})$	Normalverteilungsdichte mit Mittelwertvektor $\underline{\mu}$ und Kovarianzmatrix $\underline{K}$
$\exp(x)$	Exponentialfunktion
$\nabla_{\underline{a}}$	Gradientenoperator bezüglich Vektor $\underline{a}$
$\underline{A}_t$	Transponierte einer Matrix $\underline{A}$
$\underline{a}_t$	transponierter Vektor $\underline{a}$
$\text{Sp}(\underline{A})$	Spur der Matrix $\underline{A}$
$ \underline{A} $	Determinante der Matrix $\underline{A}$
$\ \dots\ $	Norm
$\underline{I}_n$	Einheitsmatrix der Ordnung n
$\underline{O}$	Nullmatrix

Lebenslauf

Geburtsdatum: 29. März 1950

Geburtsort: Gunzenhausen/Mfr.

Eltern: Kurt Jäpel, kfm. Angestellter
Anna Jäpel, geb. Loy

1956-1960: Besuch der Volksschule in Nürnberg

1960-1969: Besuch des Martin-Behaim-Gymnasiums
in Nürnberg, 1969 Abitur

1969-1970: Wehrdienst

1970-1976: Studium an der Universität Erlangen
im Studienfach Informatik
1976 Abschluß mit dem Diplom

Seit 1976: Mitarbeiter am Lehrstuhl für
Informatik 5 (Mustererkennung)
der Universität Erlangen

